Backloading in the Sequential Lineup Prevents Within-Lineup Criterion Shifts That Undermine Eyewitness Identification Performance

Ruth Horry, Matthew A. Palmer, and Neil Brewer
Flinders University

Although the sequential lineup has been proposed as a means of protecting innocent suspects from mistaken identification, little is known about the importance of various aspects of the procedure. One potentially important detail is that witnesses should not know how many people are in the lineup. This is sometimes achieved by backloading the lineup so that witnesses believe that the lineup includes more photographs than it actually does. This study aimed to investigate the effect of backloading on witness decision making. A large sample (N = 833) of community-dwelling adults viewed a live “culprit” and then saw a target-present or target-absent sequential lineup. All lineups included 6 individuals, but the participants were told that the lineup included 6 photographs (nonbackloaded condition) or that the lineup included 12 or 30 photographs (backloaded conditions). The suspect either appeared early (Position 2) or late (Position 6) in the lineup. Innocent suspects placed in Position 6 were chosen more frequently by participants in the nonbackloaded condition than in either backloaded condition. Additionally, when the lineup was not backloaded, foil identification rates increased from Positions 3 to 5, suggesting a gradually shifting response criterion. The results suggest that backloading encourages participants to adopt a more conservative response criterion, and it reduces or eliminates the tendency for the criterion to become more lenient over the course of the lineup. The results underscore the absolute importance of ensuring that witnesses who view sequential lineups are unaware of the number of individuals to be seen.

Keywords: eyewitness identification, sequential lineup, response bias, backloading

Eyewitness identification tests have been an integral component of criminal investigations for many years. The sequential lineup was developed by Lindsay and Wells (1985) as a means of reducing false identifications of innocent suspects. Although the sequential lineup has taken various forms when used by police forces (see, e.g., Klobuchar, Steblay, & Caligiuri, 2006; Wells, Steblay, & Dysart, 2011), the key differences between it and the much more commonly used simultaneous lineup are (a) the lineup members are shown sequentially; (b) the witness makes a yes/no (or “not sure”) decision for each lineup member and cannot return to any previously rejected lineup member; and (c) the number of lineup members is unknown to the witness.1 It is the latter component of the sequential lineup that forms the basis of this research.

Different techniques have been used, both in the laboratory and in the field, to keep witnesses unaware of the number of photographs to be seen in a sequential lineup. These techniques include informing participants to expect a specific but incorrect (and larger) number of individuals than will be shown (e.g., Carlson, Gronlund, & Clark, 2008), providing additional spaces on a response sheet (e.g., Lindsay, Lea, & Fullord, 1991), and stacking a physical deck or booklet of photographs with additional blank filler cards (e.g., Memon & Gabbert, 2003). For the purposes of this article, all of these techniques will be considered as alternative ways of backloading a sequential lineup, as all lead a witness to believe that they will see more images than are actually included in the lineup.

Administration of lineups by computer bypasses the need for a physical stack of photographs or a response sheet. Therefore, many computerized studies leave the number of individuals in the lineup undisclosed, stating only that a series of images will be shown (e.g., Gronlund, Carlson, Dailey, & Goodsell, 2009; Humphries, Holliday, & Flowe, 2012; Sauer, Brewer, & Wells, 2008). The “undisclosed” method is also common in many police jurisdictions that use the sequential lineup and was implemented in a recent North American field experiment (see preliminary report by Wells et al., 2011). For our purposes, we also consider the undisclosed method as a form of backloading, as it ensures the witness does not

---

1 Lindsay and Wells (1985) also stipulated that the sequential lineup should be administered double blind, such that the lineup administrator is unaware of the position of the suspect in the lineup. This prevents unintentional (or intentional) cueing of the witness. However, it is now widely accepted that all lineups should be presented double blind, whether sequential or simultaneous.
know the number of images in the lineup. In the discussion, we will consider how the undisclosed method may or may not lead to different assumptions by the witness compared with other methods of backloading.

The sequential lineup procedure has received a substantial amount of attention in academic and legal settings. It continues to stimulate research (e.g., Steblay, Dietrich, Ryan, Raczynski, & James, 2011), has generated extensive policy debate (e.g., Lindsay, Mansour, Beaudry, Leach, & Bertrand, 2009; Malpass, 2006), has been investigated in field trials in collaboration with police forces (e.g., Wells et al., 2011), and has received considerable attention in the popular media (e.g., Schwartz, 2011). It is, therefore, surprising that little research has investigated the role of the individual components of the sequential procedure and the effect that these components have on eyewitness identification decisions (Malpass, Tredoux, & McQuiston-Surrett, 2009).

This research focused on the backloading component of the sequential procedure. Because no published study has systematically manipulated this factor, it is not known what effects—if any—backloading has on eyewitness identification decisions. We addressed several questions: Does backloading affect the placement of witnesses’ decision criteria in the sequential lineup? Does the amount of backloading matter, or does backloading behave in an all-or-nothing manner? Is the position of the suspect in the lineup important in backloaded and nonbackloaded sequential lineups? Does backloading affect the confidence-accuracy relationship for identification decisions? We outline our predictions for each of these questions below.

Criterion Setting in the Sequential Lineup

Most research on sequential lineups has focused on comparing accuracy rates between simultaneous and sequential lineups (see Steblay, Dysart, & Wells, 2011, for a meta-analysis). In comparison to simultaneous lineups, sequential lineups reduce false identifications of innocent suspects, with a smaller reduction in correct identifications of guilty suspects. These findings are consistent with a criterion-shift account of the simultaneous-sequential difference, wherein witnesses are less willing to choose from sequential lineups than from simultaneous lineups (Palmer & Brewer, 2012; though see Steblay, Dysart et al., 2011, for an alternative view). Note we use the term simultaneous-sequential difference instead of the sequential-superiority effect (Steblay, Dysart, Fulero, & Lindsay, 2001) because it reflects the view that sequential and simultaneous lineups may each be superior to the other in different ways and under different circumstances (Gronlund et al., 2009).

The inspiration for the sequential lineup was Wells’s (1984) relative-absolute judgment model of witness decision making. According to this model, when making relative judgments, witnesses compare the lineup members to find the best match to their memorial representation. When making absolute judgments, each lineup member is individually assessed in relation to a decision criterion. If the evidence of match exceeds the criterion, the lineup member is identified; if not, the lineup member is rejected. Relative judgments are considered problematic because they increase the risk of an innocent suspect being identified; if the culprit is not in the lineup, one of the remaining lineup members will logically be the best available match (Wells, 1984, 1993). Consistent with this view, there is evidence that accuracy is poorer for relative than for absolute judgments under at least some conditions (Clark, Erikson, & Breneman, 2011). The sequential procedure was designed to encourage witnesses to rely more heavily on absolute judgments than on relative judgments (Lindsay & Wells, 1985).

However, recent research has questioned the relative-absolute account of the simultaneous-sequential difference. Clark and Davey (2005) found that removing the target from a sequential lineup without replacement shifted choices from the target to the next-best foil. Such a target-to-foil shift had previously been demonstrated in simultaneous lineups and was thought to be compelling evidence for the use of relative judgment strategies (Wells, 1993). Further, in a recent meta-analysis, Palmer and Brewer (2012) showed that sequential lineups are characterized by a conservative criterion shift, rather than an increase in discriminability. Although this finding may be compatible with the relative-absolute judgment distinction, it would perhaps be more parsimonious to assume that similar decision strategies are used in simultaneous and sequential lineups but that witnesses demand more evidence when making positive identifications from sequential lineups. In addition, Flowe and Ebbesen (2007) found that the similarity of the foils to the target influenced the rate at which the target was chosen from sequential lineups. They argued that this effect could only be explained if witnesses mentally compared the lineup member under examination with the previously rejected lineup members stored in working memory. These studies suggest that the decision-making processes used in sequential and simultaneous lineups may be quite similar (Clark & Davey, 2005; Flowe & Ebbesen, 2007) and that the relative-absolute model may be limited in the extent to which it can explain how witnesses make decisions from sequential lineups.

If the processes that govern decision making from sequential lineups are similar to those used when viewing simultaneous lineups, how might one account for the apparent criterion shift between the two types of lineup? There may be several contributing factors, including the sequential presentation of the images in itself or the requirement of a decision for each and every photograph. Backloading may also play a major role in how stringently witnesses set their decision criterion. Consider the experience of two witnesses preparing to view a sequential lineup, with equally strong memories of the offender. Witness A is told that the lineup contains 6 individuals, but Witness B is told that the lineup contains 12 individuals. The two witnesses view the same images in the same order, and thus their experiences as the lineup unfolds are very similar. The only difference between them is that Witness B believes that there are a greater number of yet-to-be-seen photographs than Witness A. How might this affect their decisions? Witness B may be more hesitant to make a choice in case an even better match comes along later. Witness B therefore sets a more stringent decision criterion than Witness A, who believes that there are fewer images—and therefore fewer opportunities for a better match to come along. Thus, although the two witnesses may have similar ecphoric experiences as they view the lineup, Witness A will be more likely to choose than Witness B. Therefore, in regard

2 Note that our interest was not in comparing sequential lineups with simultaneous lineups, and no simultaneous condition was included in this study.
to our first and primary research question, we predicted that backloading would influence identification decisions through a criterion shift. Specifically, participants who view nonbackloaded lineups will set more lenient response criteria than participants who view backloaded lineups. This will lead to a higher choosing rate, reflected by increased “hits” in target-present lineups but also increased false picks in target-absent lineups.

Our second research question relates to how much backloading is used. Will any criterion shift between backloaded and nonbackloaded lineups be all or nothing, or will it be possible to detect a move toward even more stringent criteria as the amount of backloading increases? We can only speculate at this point, but once again, logic leads us to predict that more backloading will lead to an even more stringent response criterion. To take an extreme example, a witness who believes that a lineup will include many hundreds of photos should approach that lineup very cautiously. After all, as the lineup increases in size, the odds that any one individual is the culprit necessarily decrease. This question is interesting from a theoretical perspective, insomuch as it can inform our understanding of witness decision making. However, from an applied perspective, it may also be extremely useful to know that a witness who believes that they will see 30 photographs will or will not behave more cautiously than a witness who believes they will see 12 or 15 photos. Thus, our prediction is that backloading will cause a shift to a more conservative criterion, with the size of the criterion shift depending on how much backloading is used.

**Suspect Position and the Sequential Lineup**

Recently, some researchers have begun to explore the possibility that a witness’s decision criterion could shift over the course of a sequential lineup (e.g., Carlson et al., 2008; Gronlund et al., 2009). Goodsell, Gronlund, and Carlson (2010) attempted to fit Clark’s (2003) WITNESS model to data from several published studies that had found sequential lineup advantages. WITNESS takes into account comparisons between lineup members (relative judgments) as well as the absolute match value of the lineup member associated with the strongest evidence (absolute judgments). These two types of evidence are combined and compared with a decision threshold. A lineup member is identified if the combined sum of the relative and absolute match exceeds the criterion. Goodsell et al. (2010) found that WITNESS could not adequately model the data from sequential lineups without modifications to either the decision rule or the memory parameters of the model. For example, potential modifications to the decision rule included (a) a switch from an absolute to a relative judgment part way through a lineup, accompanied by an abrupt drop in the decision criterion (the abrupt-shift hypothesis), and (b) a gradual decrease in the decision criterion over the course of the lineup (the gradual-shift hypothesis). Both of these modifications suggest a within-lineup criterion-shift to a more lenient position, though they differ in when in the lineup the shift occurs. To tease apart these possibilities, we examined foil-choosing rates across the middle-to-late positions of the lineup. The abrupt-shift hypothesis predicts no difference in choosing rates, whereas the gradual-shift hypothesis predicts an increase in choosing rates as the lineup progresses.

If the criterion can shift over the course of a sequential lineup, how might backloading influence the magnitude of this shift? The notion that witnesses might adopt a more lenient decision criterion as a sequential lineup progresses was flagged by Lindsay and Wells (1985) in their original investigation of sequential lineups, and it was this idea that lead to the inclusion of a backloading procedure. Specifically, Lindsay and Wells backloaded the procedure “to reduce any tendency to increase the probability of making a “yes” response as the end of the pile approached” (p. 560). In extreme cases, where witnesses assume the police have arrested the culprit and their task is to select that person from the lineup, witnesses might be inclined to select the last lineup member even before they have viewed that person (e.g., “It wasn’t any of the other five, so it must be this one”). Backloading should prevent (or at least reduce) a within-lineup criterion shift, by limiting the scope for witnesses to make such inferences. Presumably prompted by this reasoning, the majority of sequential lineup studies have adopted similar procedures (see study space analysis by Malpass et al., 2008). However, no published study has systematically manipulated backloading, therefore the logic remains untested.

This leads us to our third research question: How might the position of the suspect in the lineup influence responding in backloaded and nonbackloaded lineups? Position effects in sequential lineups received little attention until recently. Initial studies (e.g., Lindsay & Wells, 1985) found no position effects, which led to many researchers fixing the position of the suspect in sequential lineups, usually in the latter half of the lineup (e.g., Cutler & Penrod, 1988; Lindsay et al., 1991; Memon & Gabbert, 2003). However, interest in suspect position effects has been revived recently. For example, Clark and Davey (2005) manipulated the position of the suspect in relation to the next best foil, showing that a sufficiently similar foil that precedes the target will draw some choices away from the target. Thus, a target appearing late in a lineup may have less chance of being chosen, simply by virtue of some proportion of witnesses already having “spent” their choices on a plausible foil. Carlson et al. (2008) and Gronlund et al. (2009) found that guilty suspects were more likely to be chosen when placed later in the lineup but that innocent suspects were chosen with roughly equal frequency across all positions in the lineup. This led those authors to conclude that a witness’s ability to discriminate the guilty target from the foils might increase over the course of a sequential lineup (Gronlund et al., 2009).

The above studies show that suspect position may play an important, albeit complex, role in the sequential lineup. However, all of these studies used backloaded lineups. In the context of the present research, we expected position effects to be larger in nonbackloaded lineups than in backloaded lineups. As outlined earlier, Lindsay and Wells (1985) speculated that witnesses would become more willing to choose as they reached the end of a nonbackloaded sequential lineup, increasing the chances of a suspect being chosen if that suspect appears late in the lineup. However, if the witness is unaware of the number of to-be-seen images and, thus, their decision criterion is maintained throughout the lineup, then suspect position effects should be much smaller in backloaded lineups. Crucially, any change in the decision criterion should influence false identifications as well as correct identifications. As a consequence, the probabilive value of suspect identifications (i.e., the likelihood of guilt given that the suspect was identified) should be lower for nonbackloaded lineups than backloaded lineups, particularly when the suspect is placed later in the lineup.
The Confidence–Accuracy Relationship

If a witness identifies a suspect, how likely is it that the suspect is, in fact, the culprit? Investigators are routinely faced with this question and may request a statement of confidence from the witness to assess the reliability of the identification. It is not surprising, then, that the confidence–accuracy relationship has received a lot of attention in eyewitness identification research. Sporer, Penrod, Read, and Cutler (1995) conducted a meta-analysis of studies that had used the point-biserial correlation to investigate the confidence–accuracy relationship. By splitting participants into choosers and nonchoosers, they were able to show that the relationship is reasonably strong when a positive choice is made from a lineup. When a lineup is rejected, however, confidence is not meaningfully related to accuracy. Yet there remained a widespread belief among experts that confidence is not a good indicator of eyewitness accuracy (Kassin, Tubb, Hosch, & Memon, 2001).

However, an alternative approach to assessing the confidence–accuracy relationship—known as the calibration approach—has been used recently in eyewitness identification research. Studies that use this approach have documented strong confidence–accuracy relationships across a range of situations, at least for participants who choose from a lineup (e.g., Brewer, Keast, & Rishworth, 2002; Brewer & Wells, 2006; Juslin, Olsson, & Winman, 1996; Palmer, Brewer, Weber, & Nagesh, 2012; Sauer, Brewer, Zweck, & Weber, 2010; Sauerland & Sporer, 2009). Calibration involves examining the objective probability of accuracy at each level of confidence. Perfect calibration would be found if witnesses who were 100% confident were 100% accurate, witnesses who were 90% confident were 90% accurate, and so on. Using this technique, the confidence–accuracy relationship has been found to be strongest for adult witnesses who make a positive choice and who assign confidence ratings of more than 50%. However, the relationship is not perfect. Several studies have found that witnesses tend to be somewhat overconfident, with accuracy rates of around 75–90% for witnesses who report 90–100% confidence (e.g., Brewer & Wells, 2006).

To assess the confidence–accuracy relationship, we constructed a calibration curve, in which confidence is plotted against accuracy. Visual inspection of the curve can provide crucial information regarding the correspondence between confidence and accuracy over the full range of confidence ratings. Additionally, three statistics are calculated. The calibration statistic (C) varies from 0 (perfect calibration) to 1. The over/underconfidence (O/U) statistic varies from −1 (underconfidence) to +1 (overconfidence). The normalized resolution index (NRI) indicates how well confidence ratings discriminate between correct and incorrect responses, and ranges from 0 (no discrimination) to 1 (perfect discrimination). Full details of the calculation procedures are given in Brewer and Wells (2006).

Our final major research focus involved applying the confidence–accuracy calibration approach within the context of sequential lineup administration. There are no published data on this issue, most likely because of the large number of participants required to produce stable data (at least 200 per cell; Juslin et al., 1996). Here, we asked whether the confidence–accuracy relationship for sequential lineups is affected by backloading. We predicted that it will be, for two main reasons. First, conditions that reduce false choosing rates also reduce overconfidence, because O/U is closely tied to diagnosticity (Brewer & Wells, 2006). The degree of overconfidence expressed at any given point on the confidence–accuracy curve will depend on the ratio of correct-to-incorrect choices: the higher the ratio, the smaller the degree of overconfidence. Brewer and Wells demonstrated this by randomly selecting subsamples of cases that produced different target-absent base rates. Lower target-absent base rates were, of course, associated with fewer false picks. This, in turn, increased the ratio of correct-to-incorrect choices along the confidence–accuracy curve, decreasing overconfidence. Thus, if backloading reduces false choosing rates, it should also decrease overconfidence.

Second, backloading may affect participants’ metacognitive beliefs about their accuracy. A witness who identifies someone knowing that there are still many more photographs to come may think much more carefully about the likelihood that their response is correct. Reflecting on one’s identification decision has been shown to improve calibration and reduce overconfidence (Brewer et al., 2002). This reasoning also led us to predict that backloading would reduce overconfidence.

Summary

In this study, we manipulated backloading and suspect position in target-present and target-absent sequential lineups. Our aims were twofold: To answer some of the applied questions concerning the role of backloading in the sequential lineup and to advance understanding of the decision-making processes associated with the sequential lineup. We made four main predictions: (a) backloading would be associated with a conservative criterion shift; (b) the size of the shift would depend on how much backloading was used; (c) suspect position effects would be larger in nonbackloaded than in backloaded lineups; and (d) backloading would reduce the overconfidence expressed by witnesses.

Method

Participants and Design

A large sample of 833 community-dwelling adults was recruited from public places around the city of Adelaide, South Australia. Participants who reported that they were not wearing their usual corrective eyewear were excluded from the analyses (N = 77). This did not affect the patterns of results or conclusions. Of the remaining 756 participants, 394 were female; 18 participants did not provide information about their gender. Of the 623 participants who provided their age, the mean age was 37 (range = 18 to 86 years). The sample included 38 participants (approximately 5%) over 70 years of age.

A 2 (target presence: target-present, target-absent) × 2 (suspect position: 2, 6) × 3 (backloading: none, 6 photos, 24 photos) between-groups design was used, with participants randomly allocated to one of the 12 cells of the design (the n per cell ranged from 52 to 71).

Materials

Eleven undergraduate students (9 female; 10 of Caucasian or Mediterranean appearance, 1 Asian) volunteered to assist with the
experiment as “culprits.” A head-and-shoulders photograph of each culprit was taken several weeks prior to beginning data collection. Culprits were instructed not to wear the same clothing during data collection that they had worn in their lineup photographs.

For each culprit, a modal description was generated from the responses of three participants who took no part in any other aspect of the study. Features that were mentioned by at least two of the three participants were included in the description. Using the modal descriptions, a pool of potential foils was selected for each culprit from a large, publically accessible database of images. The pool for each culprit was narrowed down to 6 after visual inspection by the first and second authors, based on an assessment of general similarity to the culprit.

The fairness of the lineups was assessed in three steps, with separate groups of participants. First, eight experimenters produced descriptions of each culprit, which were combined into modal descriptions (these tended to include slightly more detail than the descriptions used to create the lineups because of the larger group of witnesses providing the description). In the second step, we assessed the suitability of each of the foils. A group of mock witnesses ($N = 12$) was presented with 11 slides, each containing the modal description and foils for one culprit. Witnesses were asked to select from each slide any faces that matched the description provided (see Palmer, Brewer, McKinnon, & Weber, 2010, for use of a similar procedure). The mean number of foils selected per slide was 4.3 out of 6 ($M = 4.3$, range = 3.4 to 5.3). At the individual foil level, 59 of the 66 foils were selected by at least 50% of witnesses ($M = 72\%$, range = 33–100%).

In the third step, we assessed the functional size of the lineup (i.e., the number of plausible lineup members). We provided 28 mock witnesses with a description of each target and the corresponding target-present or target-absent lineup. For each lineup, 14 witnesses saw the target-present version and 14 saw the target-absent version. Tredoux’s $E$ was calculated for each target-present and target-absent lineup, which is a measure of how many members of the lineup are appropriate based on the distribution of mock witness choices across all lineup members (Tredoux, 1999). The average $E$ estimate was 3.69 ($SD = 0.94$; range = 2.51–5.16) for target-present lineups and 3.75 ($SD = 0.80$; range = 2.97–5.44) for target-absent lineups. It is important that the target-absent lineups were not systematically more or less fair than the target-present lineups, $t(20) = 0.16$, $p = .87$. Across all lineups, culprits were chosen by 27.92% of witnesses, and innocent suspects were chosen by 19.48% of witnesses. Once again, target-present and target-absent lineups did not differ in the rates at which suspects were chosen, $t(20) = 1.24$, $p = .23$.

For each culprit, one image was designated as the innocent suspect on the basis of their similarity to the culprit (as judged by the authors). The remaining images served as foils, creating 6-person target-present and target-absent lineups for each culprit. Each image was scaled to 600 × 700 mm and was printed on a laminated card 1050 mm × 1500 mm in size. Blank cards were laminated for the backloaded conditions.

Procedure

The student volunteers worked in pairs, swapping roles as experimenters and culprits over the course of data collection. The experimenter’s role was to approach potential participants and conduct the identification test. Each participant received information about their rights and provided informed consent for participating in the study. The experimenter then directed the participant’s attention to a location 10 m away. The culprit, who had remained out of sight until now, then stepped into view for 10 s before stepping back out of view. The experimenter then read aloud the following instructions to the participant:

I’d now like you to try to identify the person you saw out of a group of [6/12/30] photographs. The person may or may not be present in the photos. I will show you the photos one at a time. You can take as long as you like to look at each photo, but you will only be allowed to see each photo once. For each photo, I would like you to indicate whether that is the person you saw earlier. If you respond “yes” to a photo, you will not be able to change that decision, and you will not be allowed to respond “yes” to any later photos.

After confirming that the participants understood the instructions, the experimenter removed the cards from the envelope. The experimenter held the stack of cards away from the body at chest height, so that he or she could remain unaware of the location of the suspect in the lineup. As each photo was held up, the experimenter asked, “Is it number . . . ?” If the participants responded “no,” the photograph was moved to the back of the stack, revealing the next photograph. Following a “yes” response, the experimenter asked the participants to rate their confidence on a scale ranging from 0% to 100%. If the participants rejected all six images, they were asked to rate their confidence that the culprit was not present on a 0% to 100% scale. All of the experimenters were trained to remain unaware of the location of the suspect until after the confidence statement had been recorded. Following the lineup, participants were thanked and debriefed.

Experimental Manipulations

Before each participant was approached, the culprit was responsible for preparing the lineup. The five foils were shuffled randomly, and the photo of the culprit or innocent suspect was placed in position two or six. The appropriate number of blank filler cards (0, 6, or 24) was added to the lineup, and all of the cards were placed in an envelope, which was handed to the experimenter. This protocol meant that the experimenter was unaware of the location of the suspect and as to whether the lineup was target-present or target-absent, until the lineup ended and the confidence judgment was recorded.

Results

Overview

Table 1 shows the percentages of suspect identifications, foil identifications, and lineup rejections, broken down by target present.

\[3\] Each slide also included four extra faces that differed from the modal description in terms of one characteristic, such as hair color or age. Responses to these faces were not analyzed, and none were used in the experiment. The additional faces were included because, in previous research, we have found that participants approach this task very conservatively, very rarely endorsing all of the foils in any lineup.
Table 1
Identification Decisions (%) by Backloading and Suspect Position

<table>
<thead>
<tr>
<th>Suspect position</th>
<th>Target present</th>
<th>Lineup rejection</th>
<th>Target absent</th>
<th>Lineup rejection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Suspect ID</td>
<td>Foil ID</td>
<td></td>
<td>Suspect ID</td>
</tr>
<tr>
<td>No backloading</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position 2</td>
<td>61.76</td>
<td>13.24</td>
<td>25.00</td>
<td>7.35</td>
</tr>
<tr>
<td>Position 6</td>
<td>50.70</td>
<td>33.80</td>
<td>15.50</td>
<td>36.54</td>
</tr>
<tr>
<td>6 photos</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position 2</td>
<td>42.42</td>
<td>15.15</td>
<td>42.43</td>
<td>10.91</td>
</tr>
<tr>
<td>Position 6</td>
<td>44.26</td>
<td>21.31</td>
<td>34.43</td>
<td>12.68</td>
</tr>
<tr>
<td>24 photos</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position 2</td>
<td>40.30</td>
<td>4.48</td>
<td>55.22</td>
<td>5.66</td>
</tr>
<tr>
<td>Position 6</td>
<td>40.32</td>
<td>17.74</td>
<td>41.94</td>
<td>6.45</td>
</tr>
</tbody>
</table>

Note. ID = identification.

eence, suspect position, and backloading condition. To address our first two research questions, we tested the effects of backloading and suspect position on overall patterns of discriminability (witnesses’ ability to distinguish culprits from foils) and response bias (witnesses’ tendency to choose from or reject a lineup). To address our third research question concerning position effects, we examined the data using multilevel logistic regression. Next, we used calibration analyses to assess the effects of backloading on the confidence–accuracy relationship. Finally, we assessed the effect of backloading on the probative value of suspect identifications from backloaded and nonbackloaded lineups.

For some data, the key descriptive statistics were a single value for each condition with no indication of the variance around that value (e.g., estimates of discriminability and response bias, and the C, O/U, and NRI statistics associated with calibration analysis). Because the usual parametric analyses could not be conducted on these data, inferential comparisons were made following the method used by Palmer, Brewer, and Weber (2010). A jackknifing procedure was used to estimate the standard error for each statistic (Koriat, Lichtenstein, & Fischhoff, 1980; Mosteller & Tukey, 1968; Weber & Brewer, 2006). These estimated standard errors were then used to calculate inferential 95% confidence intervals (CIs) around each statistic (Tryon, 2001). Nonoverlapping values indicate a difference is significant at the .05 level. An average E value (a parameter that reflects equivalence of standard errors) was used to facilitate multiple comparisons between conditions where appropriate (e.g., between the three backloading conditions).

Effects of Backloading on Decision Criteria

We estimated discriminability ($d'$) and response bias ($c$) using a recent signal detection-based model called signal detection theory–compound decisions (SDT-CD; Duncan, 2006). This model is designed to describe decision making in compound tasks such as eyewitness identification, in which a participant must decide whether an item is present in a series and, if so, which item is the correct item (for a description of the model, see Palmer & Brewer, 2012; Palmer, Brewer, & Weber, 2010). SDT-CD estimates of discriminability ($d'$) and response bias ($c$) take into account all choices from target-present and target-absent lineups—whether of foils, targets, or innocent suspects, and therefore provide a comprehensive treatment of the data. Conceptually, the estimates of discriminability and bias produced by SDT-CD can be interpreted the same way as those produced by simpler signal detection models. For $d'$, increasing positive values indicate greater ability to distinguish culprits from foils, and a value of 0 indicates chance performance. For $c$, positive and negative values indicate conservative and lenient responding, respectively, with 0 representing unbiased responding (the point at which accuracy is maximized when lineups are equally likely to be target present or target absent).

For discriminability, the CIs of the backloading conditions were overlapping. This suggests that backloading did not influence discriminability. However, inspection of Figure 1a shows that for nonbackloaded lineups, discriminability was higher when the suspect was in Position 2, $d' = 2.28$, 95% CI = 1.90, 2.66, than when the suspect was in Position 6, $d' = 1.49$, 95% CI = 1.26, 1.72. We will discuss this point later when we discuss position effects.

For response bias, the results were much more striking (see Figure 1b). Across the sample, responding was more lenient in the nonbackloaded condition than either of the backloaded conditions. This supports our first hypothesis, that backloading would operate through a criterion shift. Our second research question was whether the amount of backloading was important. When the data were split by suspect position, the differences between the two backloaded conditions were not significant. However, to increase our statistical power, we collapsed the data across the two suspect position conditions. When we did, the results showed that the participants were more conservative in the high-backloading condition ($c = -0.19$, 95% CI = $-0.31$, $-0.07$) than in the moderate-backloading condition ($c = -0.95$, 95% CI = $-1.16$, $-0.74$). This supports our second prediction, that responding would become more conservative as the amount of backloading increased. In addition, although responding was lenient in all conditions (as evidenced by negative $c$ values), responding became less biased as backloading increased (evidenced by $c$ values closer to zero).

Suspect Position Effects

Our third research question concerned position effects in backloaded and nonbackloaded lineups. We used a logistic regression model to predict identification responses from suspect position, backloading condition, and target presence. We used a multilevel
model because we had 11 different culprits, who all contributed a different number of trials to the analysis (see Wright & London, 2009). Furthermore, the patterns of identification responses varied across the culprits (see Table 2). The variation between culprits was statistically significant for suspect identifications, foil identifications, and lineup rejections for target-present lineups and foil identifications for target-absent lineups (all Cramer’s Vs > .23, ps < .05).

We nested trials within culprits to account for the shared variance between trials from the same culprit (for a similar treatment of eyewitness identification data see Horry, Memon, Wright, & Milne, 2012). The data were analyzed using the lme4 package for R (Bates, Maechler, & Bolker, 2011). We used a backward elimination method, starting with a model that contained all main effects and interactions. At each step, the effect with the smallest predictive value (i.e., the highest p value) was removed, beginning with higher order interactions. Effects contained within high-order interactions could not be removed (cf. Wright, Gabbert, Memon, & London, 2008). Effect sizes are reported as the natural log of the odds ratio (lnOR). A lnOR of 0 indicates no difference between the two conditions, whereas a positive lnOR indicates a difference in the hypothesized direction.

The analysis of suspect identifications (see Table 1) revealed a significant three-way interaction between target presence, backloading, and suspect position, \( \chi^2(2, N = 756) = 8.57, p = .01 \). To understand this interaction, target-present and target-absent lineups were analyzed separately.

For target-present lineups, the main effect of backloading was significant, \( \chi^2(2, N = 395) = 8.62, p = .01 \). Culprits were identified more frequently from nonbackloaded lineups (56.12%) than from lineups backloaded with either 6 photos (43.31%, \( z = 2.24, p = .03 \), lnOR = 0.26) or 24 photos (40.31%, \( z = 2.76, p = .006 \), lnOR = 0.33). Again, these data are consistent with our predictions concerning a criterion shift. The two backloaded conditions did not significantly differ, \( z = 0.51, p = .61 \), lnOR = 0.07. The main effect of suspect position was not significant, \( \chi^2(1, N = 395) = 0.71, p = .40 \), lnOR = −0.06, and the interaction term did not approach significance, \( \chi^2(2, N = 395) = 1.49, p = .48 \). Thus, for target-present lineups, position effects were not detected, even for nonbackloaded lineups.

For target-absent lineups, the main effect of suspect position was significant, \( \chi^2(1, N = 361) = 9.86, p = .002 \), lnOR = 0.78. Innocent suspects in Position 6 (17.30%) were more likely to be chosen innocent suspects in Position 2 (7.95%). The main effect of backloading was also significant, \( \chi^2(2, N = 361) = 10.39, p = .006 \). Innocent suspects were more likely to be chosen from nonbackloaded lineups (20.00%) than from lineups backloaded with 24 photos (6.09%, \( z = 2.95, p = .003 \), lnOR = 1.19). The difference between nonbackloaded lineups and lineups with 6 extra photos (11.90%) approached significance, \( z = 1.73, p = .08 \), lnOR = 0.52. However, all of the above main effects should be interpreted in light of a significant interaction between suspect position and backloading, \( \chi^2(2, N = 361) = 6.89, p = .03 \). For target-absent lineups that were not backloaded, innocent suspects in Position 6 were more likely to be chosen (36.54%) than innocent suspects in Position 2 (7.35%). \( \chi^2(1, N = 120) = 16.68, p < .001 \), lnOR = 1.60. However, no position effects were observed for lineups backloaded with 6 photos, \( \chi^2(1, N = 126) = 0.09, p = .76 \), lnOR = 0.15; or 24 photos, \( \chi^2(1, N = 115) = 0.04, p = .84 \), lnOR = 0.13. These results support our predictions that suspect position effects would be larger for nonbackloaded lineups than for backloaded lineups. The magnitude of the effect in the nonbackloaded condition is quite striking: Innocent suspects in Position 6 were almost five times as likely to be chosen as the same suspects placed in Position 2. These results also suggest that backloading becomes increasingly important as the lineup progresses. In other words, the effects of backloading are indeed moderated by the position of the suspect in the lineup. The above results also explain why suspect position influenced discriminability in our SDT-CD analysis. Suspect position influenced false choosing rates but not correct suspect identifications. Consequently, the discriminability estimate was higher when the suspect appeared earlier, compared with later, in the lineup.

The above analyses showed that innocent suspects were more likely to be identified if they were placed in Position 6 rather than in Position 2 but that culprits were chosen with similar frequency regardless of their position in the lineup. However, in a sequential lineup, the available pool of witnesses who could potentially
identify the suspect decreases as the lineup progresses. After all, witnesses who identify a foil prior to the appearance of the suspect have already “spent their pick” leaving them with no opportunity to identify the suspect (Clark & Davey, 2005). Failing to account for prior-to-suspect foil identifications may obscure any influence of choosing tendencies.

Following Clark and Davey (2005), we conditionalized suspect identification rates on the number of witnesses who had not already identified a foil. To clarify the rationale for this approach, consider the following example. One hundred witnesses see a six-person lineup in which the suspect appears in Position 6. The suspect is identified by 25 of the witnesses, giving a suspect identification rate of 25%. However, imagine that each preceding foil was chosen by 10 witnesses. As a result, 50 of the witnesses had already made a choice and were unable to choose the suspect. If we remove these witnesses from the equation, the conditionalized suspect rate becomes 25/50, or 50%. Because the number of “spent” witnesses can only increase or remain the same over the course of the lineup, conditionalized and unconditionalized suspect identification rates will usually diverge more as the lineup progresses.

Figures 2a and 2b show the conditionalized suspect identification rates in target-present and target-absent lineups, respectively, broken down by backloading condition and suspect position. Chi-square analyses showed that innocent suspects in nonbackloaded lineups were more likely to be chosen in Position 6 (48.72%) than in Position 2 (7.69%), $\chi^2(1) = 23.11$, $p < .001$, $ln OR = 1.84$. There was no evidence of position effects in any other cell of the design (all $\chi^2$s $\leq 1.85$, $p > .10$). These results converge with those of the nonconditionalized suspect identification rates. Taken together, the results suggest that participants viewing a nonbackloaded lineup adjusted their decision criterion over the course of the lineup but that participants in the backloaded conditions did not (or did so at a rate that could not be detected given this sample size).

How might the decision criterion change over the course of a nonbackloaded lineup? One possibility is that the change is gradual, with the criterion becoming progressively more lenient after each member of the lineup is eliminated. Another possibility is that participants abruptly drop their criterion for the final lineup member. To test these two possibilities, we examined foil choices from nonbackloaded lineups (collapsed across target-presence and suspect position) across positions 3, 4, and 5, conditionalized on witnesses who had not already made a prior choice (see Table 3). We did not include Position 2 because, for half of the participants, Position 2 was occupied by the suspect. The chi-square test showed that the proportion of foil identifications differed across the three lineup positions, $\chi^2(2) = 6.95$, $p = .03$. Foil identification rates increased from 3.38% in Position 3 to 7.00% in Position 4, to 13.44% in Position 5. These results are consistent with a gradually shifting decision criterion in a nonbackloaded lineup.

For completeness, we also analyzed lineup rejections and foil identifications by using the multilevel logistic regression method. The analysis of lineup rejections revealed no significant interactions. However, all three main effects were significant. Lineups without backloading were rejected less frequently (34.36%) than lineups backloaded with either 6 (50.99%, $z = 3.90$, $p < .001$, $ln OR = 0.39$) or 24 (65.16%, $z = 7.09$, $p < .001$, $ln OR = 0.64$) photos, $\chi^2(2, N = 756) = 53.81$, $p < .001$. The difference between 6 and 24 photos was also significant, $z = 3.46$, $p < .001$, $ln OR = 0.25$. Lineups were rejected more frequently when the suspect was in Position 2 (53.58%) than when the suspect was in Position 6 (46.17%), $\chi^2(1, N = 756) = 5.10$, $p = .02$, $ln OR = 0.15$. Target-absent lineups (65.65%) were rejected more frequently than target-present lineups (34.55%), $\chi^2(1, N = 756) = 70.86$, $p < .001$, $ln OR = 0.64$.

The analysis of foil identifications revealed a significant two-way interaction between target-presence and suspect position, $\chi^2(1, N = 756) = 10.61$, $p = .001$. In target-present lineups, foil choices were more frequent when the suspect appeared later in the lineup (24.74%) than when the suspect appeared early in the lineup (10.95%), $\chi^2(1, N = 395) = 14.63$, $p < .001$, $ln OR = 0.82$. In target-absent lineups, foil identifications were not associated with suspect position, $\chi^2(1, N = 361) = 1.09$, $p = .30$. There was also...
a significant main effect of backloading, $\chi^2(2, N = 756) = 21.75, p < .001$. Lineups backloaded with 24 photos produced fewer foil identifications (10.85%) than nonbackloaded lineups (23.74%), $z = 4.39, p = .001, lnOR = -0.78$, or lineups backloaded with 6 photos (18.11%), $z = 3.21, p = .001, lnOR = -0.51$. Again, this is consistent with a criterion shift as the amount of backloading increases.

### Table 3

<table>
<thead>
<tr>
<th>Condition</th>
<th>Lineup position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Not backloaded</td>
<td>3.38% (7/207)</td>
</tr>
<tr>
<td>Backloaded</td>
<td>1.66% (7/421)</td>
</tr>
</tbody>
</table>

*Note.* Both backloading conditions have been combined to increase cell frequencies. Frequencies are shown in parentheses.

The Confidence–Accuracy Relationship

Large numbers of data points are required (at least 200 per cell of the design, Juslin et al., 1996) to create stable calibration curves. To allow for sufficient data points, we collapsed across the suspect position variable, looking only at the main effect of backloading ($ns$ between 265 and 292 per condition). This allowed us to address our fourth research question: Does backloading affect the confidence–accuracy relationship? Following prior research, we collapsed the confidence scale into a 5-point scale (0–20%, 20–40%, 50–60%, 78–80%, 90–100%; see Brewer & Wells, 2006; Juslin et al., 1996; Sauer et al., 2010). Figure 3 shows the calibration curves for the three backloading conditions for choosers and nonchoosers. The dotted diagonal line on each graph represents perfect calibration. Points that fall below the line indicate overconfidence, and points that fall above the line represent underconfidence. The C, O/U, and NRI statistics are shown in Table 4.

Several aspects of the confidence–accuracy data are noteworthy. First, in all backloading conditions, the degree of calibration was good for choosers but poorer for nonchoosers. This is evident in the slopes of the calibration curves (which followed the perfect calibration line quite closely for choosers but were flat for non-
choosers) and C statistics. Recall that C varies from 0 to 1, with 0 indicating perfect calibration. Table 4 shows that, for choosers, the C values were close to zero and the inferential 95% CIs included zero for the 6-photo and 24-photo conditions. For nonchoosers, the C values were somewhat higher and did not include zero in any condition. These results replicate previous findings from simultaneous lineups that confidence is better calibrated for choosers than for nonchoosers (Brewer et al., 2002; Brewer & Wells, 2006; Palmer et al., 2012; Sauerland & Sporer, 2009). For choosers, calibration did not vary significantly between the backloading conditions (evidenced by overlapping CIs). For nonchoosers, calibration was poorer in the no-backloading condition than the 24-photo condition.

Second, for choosers, backloading a lineup with 24 photos reduced overconfidence. The curve for the 24-photo condition follows the perfect calibration line very closely, whereas the curves for the no-backloading and 6-photo conditions are below the ideal line, particularly in the upper half of the confidence scale. Further, O/U, which varies from −1 (underconfidence) to +1 (overconfidence), was lower for the 24-photo condition than the nonbackloaded or 6-photo conditions (although the comparison between the 6- and 24-photo conditions only approached significance). Indeed, the O/U value for the 24-photo condition was approximately zero, indicating no tendency for choosers in this condition to be either overconfident or underconfident. (Note also that, for this condition, the SEs overlapped the perfect calibration line for all points of the curve bar the 40–60% category, which displayed slight underconfidence.) The shift from no backloading to some backloading did not, in itself, reduce overconfidence. However, in the high-backloading condition, overconfidence was eliminated. A different pattern was found for nonchoosers, who tended to be underconfident in the no-backloading condition but showed no tendency to be either over- or underconfident in the high-backloading condition. Taken together, these results suggest that leading participants to believe that there are many images yet to come may cause them to reflect more on the quality of their memory for the target, ultimately leading them to produce more realistic confidence assessments (cf. Brewer et al., 2002).

Third, backloading had no significant effect on resolution for choosers or nonchoosers (evidenced by overlapping CIs around the NRI statistics for each backloading condition). The CIs for choosers did not include 0 in any of the backloading conditions, indicating that confidence did discriminate correct from incorrect choices. The NRI can be interpreted as eta-squared, with cutoffs of .010, .059, and .138, for small, medium, and large effects, respectively (Brewer & Wells, 2006). The data therefore show that the confidence–accuracy relationship was moderate in the nonbackloaded condition and strong in both backloaded conditions.

### Probative Value

From an applied perspective, it is important to know whether a particular response makes it more likely that a suspect is guilty or innocent and, if so, by how much. This is commonly referred to as probative value or diagnosticity. There are different ways to calculate probative value, which rest on different assumptions (e.g., assumptions about the culprit-present base rate). Here, we calculate probative value as the natural log of the odds ratio of correct suspect identifications to incorrect suspect identifications. We use this particular measure because it makes no assumptions about the culprit-present base rate. Instead, it provides a measure of by how much one should adjust their belief in the guilt or innocence of the suspect following a particular type of decision. Positive values indicate that the one should be more convinced of the suspect’s guilt following the decision, while negative values indicate that one should be more convinced of the suspect’s innocence following the decision. Figure 4 shows the probative value of suspect identifications, with inferential 95% CIs.

From Figure 4, it is apparent that one cell of the design differs from all of the others—nonbackloaded lineups with the suspect in Position 6. The upperbound of the CI is lower than the lowerbound in every other condition, indicating significantly lower probative value. In fact, the CI for this condition included 0, which suggests no probative value at all. In other words, if a suspect was identified in the final position of a nonbackloaded lineup, one should be no

#### Table 4

<table>
<thead>
<tr>
<th>Backloading condition</th>
<th>Choosers</th>
<th>Nonchoosers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SE</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 photos</td>
<td>.10</td>
<td>.03</td>
</tr>
<tr>
<td>6 photos</td>
<td>.04</td>
<td>.02</td>
</tr>
<tr>
<td>24 photos</td>
<td>.02</td>
<td>.01</td>
</tr>
<tr>
<td>O/U</td>
<td>.00</td>
<td>.01</td>
</tr>
<tr>
<td>0 photos</td>
<td>.12</td>
<td>.05</td>
</tr>
<tr>
<td>6 photos</td>
<td>.14</td>
<td>.04</td>
</tr>
<tr>
<td>24 photos</td>
<td>.14</td>
<td>.09</td>
</tr>
<tr>
<td>NRI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 photos</td>
<td>.12</td>
<td>.06</td>
</tr>
<tr>
<td>6 photos</td>
<td>.14</td>
<td>.07</td>
</tr>
<tr>
<td>24 photos</td>
<td>.14</td>
<td>.09</td>
</tr>
</tbody>
</table>

**Note.** C = calibration statistic; O/U = over/underconfidence; NRI = normalized resolution index; SE = standard error; 95% CIs = 95% confidence intervals.
important to note that backloading prompted witnesses not only to reduce incorrect responses but also to set a more conservative criterion but also a less biased criterion; that is, participants shifted from being too lenient to being closer to the optimal position (i.e., zero). As a result, the reduction in correct identifications was outweighed by a larger reduction in false identifications. This result implies that backloading will increase the accuracy of identification decisions in situations where target-present and target-absent lineups are equally likely to occur (e.g., experiments). However, conservative responding does not necessarily translate to increased accuracy in all situations. For example, in situations where target-present lineups are far more common than target-absent lineups, accuracy will be maximized through lenient responding. Thus, because the base rate of target-present lineups in police investigations is not known, it cannot be assumed that backloading will always increase identification accuracy in applied settings. However, to the extent that legal systems should prioritize the protection of the innocent (e.g., Laufer, 1995) and should seek to avoid evidence based on guesses, lineup procedures that promote more conservative choosing should perhaps be favored.

Our second research question concerned whether backloading would behave in an all-or-nothing manner, or whether it would be possible to detect shifts to more and more conservative positions as the amount of backloading was increased. Our signal detection analysis seemed to confirm the second possibility, as participants who expected to see 30 photographs were more conservative than those who expected to see 12 photographs. Closer inspection of the data showed that the difference between the two backloaded conditions was mainly driven by incorrect choosing rates, whereas correct identifications were similar in the two groups. However, these differences in response criteria did not translate into meaningful differences in terms of probative value. Thus, although an increased amount of backloading produced a greater criterion shift, there were no significant benefits from an investigative perspective. Clearly, more research is needed to examine how different levels of backloading affect response bias and probative value of suspect identifications if sequential procedures are to be used most effectively in police investigations.

Our third research question was whether the position of the suspect was different in backloaded and nonbackloaded lineups. We had predicted that nonbackloaded lineups might produce within-lineup criterion shifts, which would lead to suspects placed late in a lineup being chosen more frequently than the same suspects placed earlier in the lineup. Our results confirmed these predictions. In nonbackloaded lineups, foil choices increased across Positions 3 to 5, suggesting that witnesses were gradually shifting their decision criteria to ever more lenient positions. Consequently, suspects in Position 6 of nonbackloaded lineups were chosen more frequently than suspects in Position 2, and this effect was particularly striking for innocent suspects, with around a fivefold increase in false identifications. This result implies that backloading will increase the accuracy of identification decisions in situations where target-present and target-absent lineups are equally likely to occur (e.g., experiments). However, conservative responding does not necessarily translate to increased accuracy in all situations. For example, in situations where target-present lineups are far more common than target-absent lineups, accuracy will be maximized through lenient responding. Thus, because the base rate of target-present lineups in police investigations is not known, it cannot be assumed that backloading will always increase identification accuracy in applied settings. However, to the extent that legal systems should prioritize the protection of the innocent (e.g., Laufer, 1995) and should seek to avoid evidence based on guesses, lineup procedures that promote more conservative choosing should perhaps be favored.

Our second research question concerned whether backloading would behave in an all-or-nothing manner, or whether it would be possible to detect shifts to more and more conservative positions as the amount of backloading was increased. Our signal detection analysis seemed to confirm the second possibility, as participants who expected to see 30 photographs were more conservative than those who expected to see 12 photographs. Closer inspection of the data showed that the difference between the two backloaded conditions was mainly driven by incorrect choosing rates, whereas correct identifications were similar in the two groups. However, these differences in response criteria did not translate into meaningful differences in terms of probative value. Thus, although an increased amount of backloading produced a greater criterion shift, there were no significant benefits from an investigative perspective. Clearly, more research is needed to examine how different levels of backloading affect response bias and probative value of suspect identifications if sequential procedures are to be used most effectively in police investigations.

Our third research question was whether the position of the suspect was different in backloaded and nonbackloaded lineups. We had predicted that nonbackloaded lineups might produce within-lineup criterion shifts, which would lead to suspects placed late in a lineup being chosen more frequently than the same suspects placed earlier in the lineup. Our results confirmed these predictions. In nonbackloaded lineups, foil choices increased across Positions 3 to 5, suggesting that witnesses were gradually shifting their decision criteria to ever more lenient positions. Consequently, suspects in Position 6 of nonbackloaded lineups were chosen more frequently than suspects in Position 2, and this effect was particularly striking for innocent suspects, with around a fivefold increase in false identifications. This result implies that backloading will increase the accuracy of identification decisions in situations where target-present and target-absent lineups are equally likely to occur (e.g., experiments). However, conservative responding does not necessarily translate to increased accuracy in all situations. For example, in situations where target-present lineups are far more common than target-absent lineups, accuracy will be maximized through lenient responding. Thus, because the base rate of target-present lineups in police investigations is not known, it cannot be assumed that backloading will always increase identification accuracy in applied settings. However, to the extent that legal systems should prioritize the protection of the innocent (e.g., Laufer, 1995) and should seek to avoid evidence based on guesses, lineup procedures that promote more conservative choosing should perhaps be favored.

Our second research question concerned whether backloading would behave in an all-or-nothing manner, or whether it would be possible to detect shifts to more and more conservative positions as the amount of backloading was increased. Our signal detection analysis seemed to confirm the second possibility, as participants who expected to see 30 photographs were more conservative than those who expected to see 12 photographs. Closer inspection of the data showed that the difference between the two backloaded conditions was mainly driven by incorrect choosing rates, whereas correct identifications were similar in the two groups. However, these differences in response criteria did not translate into meaningful differences in terms of probative value. Thus, although an increased amount of backloading produced a greater criterion shift, there were no significant benefits from an investigative perspective. Clearly, more research is needed to examine how different levels of backloading affect response bias and probative value of suspect identifications if sequential procedures are to be used most effectively in police investigations.

Our third research question was whether the position of the suspect was different in backloaded and nonbackloaded lineups. We had predicted that nonbackloaded lineups might produce within-lineup criterion shifts, which would lead to suspects placed late in a lineup being chosen more frequently than the same suspects placed earlier in the lineup. Our results confirmed these predictions. In nonbackloaded lineups, foil choices increased across Positions 3 to 5, suggesting that witnesses were gradually shifting their decision criteria to ever more lenient positions. Consequently, suspects in Position 6 of nonbackloaded lineups were chosen more frequently than suspects in Position 2, and this effect was particularly striking for innocent suspects, with around a fivefold increase in false identifications. This result implies that backloading will increase the accuracy of identification decisions in situations where target-present and target-absent lineups are equally likely to occur (e.g., experiments). However, conservative responding does not necessarily translate to increased accuracy in all situations. For example, in situations where target-present lineups are far more common than target-absent lineups, accuracy will be maximized through lenient responding. Thus, because the base rate of target-present lineups in police investigations is not known, it cannot be assumed that backloading will always increase identification accuracy in applied settings. However, to the extent that legal systems should prioritize the protection of the innocent (e.g., Laufer, 1995) and should seek to avoid evidence based on guesses, lineup procedures that promote more conservative choosing should perhaps be favored.

Our second research question concerned whether backloading would behave in an all-or-nothing manner, or whether it would be possible to detect shifts to more and more conservative positions as the amount of backloading was increased. Our signal detection analysis seemed to confirm the second possibility, as participants who expected to see 30 photographs were more conservative than those who expected to see 12 photographs. Closer inspection of the data showed that the difference between the two backloaded conditions was mainly driven by incorrect choosing rates, whereas correct identifications were similar in the two groups. However, these differences in response criteria did not translate into meaningful differences in terms of probative value. Thus, although an increased amount of backloading produced a greater criterion shift, there were no significant benefits from an investigative perspective. Clearly, more research is needed to examine how different levels of backloading affect response bias and probative value of suspect identifications if sequential procedures are to be used most effectively in police investigations.

Our third research question was whether the position of the suspect was different in backloaded and nonbackloaded lineups. We had predicted that nonbackloaded lineups might produce within-lineup criterion shifts, which would lead to suspects placed late in a lineup being chosen more frequently than the same suspects placed earlier in the lineup. Our results confirmed these predictions. In nonbackloaded lineups, foil choices increased across Positions 3 to 5, suggesting that witnesses were gradually shifting their decision criteria to ever more lenient positions. Consequently, suspects in Position 6 of nonbackloaded lineups were chosen more frequently than suspects in Position 2, and this effect was particularly striking for innocent suspects, with around a fivefold increase in false identifications. This result implies that backloading will increase the accuracy of identification decisions in situations where target-present and target-absent lineups are equally likely to occur (e.g., experiments). However, conservative responding does not necessarily translate to increased accuracy in all situations. For example, in situations where target-present lineups are far more common than target-absent lineups, accuracy will be maximized through lenient responding. Thus, because the base rate of target-present lineups in police investigations is not known, it cannot be assumed that backloading will always increase identification accuracy in applied settings. However, to the extent that legal systems should prioritize the protection of the innocent (e.g., Laufer, 1995) and should seek to avoid evidence based on guesses, lineup procedures that promote more conservative choosing should perhaps be favored.

Our second research question concerned whether backloading would behave in an all-or-nothing manner, or whether it would be possible to detect shifts to more and more conservative positions as the amount of backloading was increased. Our signal detection analysis seemed to confirm the second possibility, as participants who expected to see 30 photographs were more conservative than those who expected to see 12 photographs. Closer inspection of the data showed that the difference between the two backloaded conditions was mainly driven by incorrect choosing rates, whereas correct identifications were similar in the two groups. However, these differences in response criteria did not translate into meaningful differences in terms of probative value. Thus, although an increased amount of backloading produced a greater criterion shift, there were no significant benefits from an investigative perspective. Clearly, more research is needed to examine how different levels of backloading affect response bias and probative value of suspect identifications if sequential procedures are to be used most effectively in police investigations.

Figure 4. Probative value of suspect identifications with inferential 95% confidence intervals.

more convinced that the suspect is guilty than that the suspect is innocent. The CIs in all other conditions were overlapping and did not include 0, indicating that suspect identifications did not differ significantly in their probative value for these conditions.

Discussion

This research was driven by several unanswered questions regarding the role of backloading in the sequential lineup: How does backloading influence witnesses’ decision making? Does the amount of backloading make a difference? Does the position of the suspect in the lineup matter in backloaded and nonbackloaded sequential lineups? And how does backloading affect the confidence–accuracy relationship?

In regard to our first and primary research question, the data suggest that backloading influences decision making by encouraging witnesses to set a more conservative decision criterion. Participants were less willing to choose from backloaded lineups than from nonbackloaded lineups, regardless of whether the suspect appeared in Position 2 or 6. We note that the criterion shift was not accompanied by an increase in discriminability, as backloading affected both correct and incorrect choosing rates. These results make perfect sense when viewed from a signal-detection perspective. Discriminability should be affected by such factors as memory strength and the ecphoric experience when viewing the lineup (influenced by such factors as the similarity between the target and the foils). Backloading does not directly influence the underlying distributions of memory activation associated with the lineup members, and therefore we would expect that any influences on behavior would be through response bias rather than through changes in discriminability.

An inevitable consequence of changes in response criteria is that reductions in incorrect responses are accompanied by reductions in correct responses. This trade-off has led some researchers to caution against advocating the sequential procedure to policymakers (e.g., Meissner, Tredoux, Parker, & MacLin, 2005). However, it is important to note that backloading prompted witnesses not only to
Position 6 than in Position 2. These data are particularly encouraging when one considers that it is common (and recommended) practice in the field that witnesses should be unaware of the number of individuals included in a lineup. Our results suggest that if this condition is met, suspect position effects may be very small in sequential lineups (though see Carlson et al., 2008, and Gronlund et al., 2009). Consequently, the probative value of the identification will likely not vary depending on the position of the suspect in the lineup. We draw this conclusion cautiously, as we did not examine all possible lineup positions. However, the data strongly suggest that being unaware of the true number of images in a lineup encourages witnesses to both set and maintain a fairly conservative decision criterion over the course of the entire lineup. This may be especially true in the field where witnesses may be allowed to make multiple choices that can then be resolved at the end of the procedure. The somewhat arbitrary rules for categorizing multiple choices in the laboratory (e.g., first-choice counts vs. last-choice counts) may have exaggerated or even created position effects in the laboratory where they have been found.

Finally, our large sample allowed us to provide the first test of the confidence–accuracy relationship in sequential lineups using the calibration approach. The results were quite similar to previous findings from simultaneous lineups. We found reasonably good calibration for choosers—though choosers did tend to be overconfident at the top end of the confidence scale (cf. Brewer & Wells, 2006; Palmer et al., 2012; Sauer et al., 2010; Sauerland & Sporer, 2009). We had predicted that backloading might influence overconfidence, and this prediction was supported: Participants who believed that they were going to see 30 photographs showed no systematic tendency toward overconfidence at all. There are two potential explanations for this finding. On the one hand, the drop in overconfidence may be a statistical consequence of the reduction in false-identification rates in the high-backloading condition. Brewer and Wells (2006) manipulated target-absent base rates to show that conditions that produce higher diagnosticity (by reducing the number of false picks) also produce less overconfidence. An alternative explanation is that participants in the high-backloading condition were more realistic in their assessments of their likely accuracy. Brewer et al. (2002) found that witnesses who were encouraged to reflect on their identification decision, or who were asked to think about whether they might have been wrong, produced confidence assessments that were better calibrated to their accuracy. Perhaps the knowledge that there were many more faces yet to be seen encouraged participants in the high backloading to consider their decision in more depth, reflecting on the likelihood that a better match was yet to come. Clearly, further research is needed to tease apart this metacognitive explanation from the statistical one.

Limitations and Future Directions

Though this study provided the first empirical data on the effects of backloading, there are many questions that are ripe for future research. Such questions include whether backloading might interact with other forensically relevant variables. For example, the criterion shifts associated with backloading may vary with retention interval and witnesses’ realizations that their memory may be weaker than it was at the time of the crime. We also know that exposure duration and retention interval have some influence on the degree of overconfidence shown for simultaneous lineups (Palmer et al., 2012; Sauer et al., 2010), so it would not be surprising if similar patterns were found with the sequential lineup. But whether such patterns would be moderated by backloading remains an empirical question.

For ease of implementation in the field setting where the research was conducted, we used a strict stopping rule. Participants were informed that they could only make a single positive identification and that the lineup would end if they chose to identify someone. Though strict stopping rules have been used in some previous laboratory research (e.g., Kneller, Memon, & Stevenage, 2001; Memon & Gabbert, 2003; Wilcock, Bull, & Vrij, 2005), it is more common to allow participants to see all images in the lineup, regardless of any positive identifications made (see meta-analysis by Steblay et al., 2011). It is also common practice among police departments in the United States to allow witnesses to see all lineup members and to allow multiple identification decisions that can be clarified by the witness at the end of the procedure (Klobuchar et al., 2006; Wells et al., 2011). Our choice of a strict stopping rule may have influenced participants’ criterion setting. Specifically, participants may have behaved more conservatively than they would have given the opportunity to make multiple positive identifications, and this, in turn, may have exaggerated the differences between the backloading conditions. A witness who knows they can make just one choice out of 30 expected photographs may behave more cautiously than one who is told they can make more than one “yes” or “not sure” response to the same set of photographs. However, though procedures that more closely mirror police practice may produce smaller effects than those reported here, we believe that backloading would continue to operate on witness decisions through the same mechanism. Regardless of stopping rule, in the absence of backloading, participants would likely shift their decision criterion over the course of the lineup, increasing the risk to an innocent suspect who happens to appear in a late lineup position. Future research should, of course, investigate the potential moderating effect of stopping rule on the backloading effect to see whether the effects reported here generalize to procedures that are closer to current practice in the field.

Our manipulation of backloading was quite strong. Witnesses were informed explicitly at the beginning of the lineup how many images to expect to see. To ensure that the manipulation was plausible, we also adjusted the size of the physical stack of photographs being held by the experimenter. This strong manipulation allowed us the best possible chance of detecting differences between conditions that would allow us to answer our fundamental research question: How does awareness (vs. lack of awareness) of the size of a sequential lineup affect decision making? In addition, although the manipulation may seem quite heavyhanded, each element of the approach has been used in previous sequential lineup studies (e.g., Carlson et al., 2008; Lindsay et al., 1991; Memon & Gabbert, 2003).

However, it is crucial that future research conducts empirical tests of the impact on decision making of the undisclosed method, in which participants are given no indication of how many images to expect. This method is common in laboratory studies that use computerized lineup administration, as there is no need for a physical response form or stack of photos (e.g., Gronlund et al., 2009; Humphries et al., 2012; Sauer et al., 2008). The undisclosed
method is also in common usage by law enforcement agencies that have adopted the sequential procedure (see Wells et al., 2011). From an applied perspective, the undisclosed method allows witnesses to remain unaware of the number of images in a lineup without any need to deceive (either directly or indirectly) the witness. However, without empirical data, it is impossible to know how the undisclosed method affects witnesses’ decision criteria. One possibility is that the results would look quite similar to those found here—that witnesses would set and maintain strict decision criteria in the face of such uncertainty. However, it is also possible that witnesses could make inferences about the typical six-pack lineup, familiar to many people through the media. Such an inference could lead to a similar within-lineup criterion-shift as that seen in the nonbackloaded condition of this study. Only an empirical test of the undisclosed method will be able to shed any light on these possibilities.

Our data also raise another question: To what degree is backloading responsible for the simultaneous–sequential difference found in the laboratory? A recent meta-analysis by Palmer and Brewer (2012) showed that the simultaneous–sequential difference is due to more conservative responding for sequential versus simultaneous lineups. Could the more conservative responding found for sequential lineups have been due to backloading? In 20 of the 22 sequential lineups analyzed by Palmer and Brewer, the witnesses were unaware of the number of images to be seen; it is not surprising that this was not the case with any of the simultaneous lineups. It remains an open question as to whether (or by how much) the sequential–simultaneous difference would disappear if sequential lineups were not backloaded, or if simultaneous and sequential lineups were all backloaded (Malpass et al., 2008; McQuiston-Surrett, Malpass, & Tredoux, 2006). For example, researchers could backload a simultaneous lineup by informing participants that they will see photographs in groups of 6 (or 8, 10, 12, etc.) and that the order in which they will see the groups is randomly determined. In reality, however, the lineup would consist of only one group (which would therefore contain the culprit or innocent suspect), but the witnesses would be led to believe that there were more images to be seen and that the suspect may be included among later groups of images. Of course, such a lineup would no longer be a strictly simultaneous lineup. Rather, it would be a hybrid simultaneous–sequential lineup, incorporating elements of each procedure.

Conclusion

This study provides the first empirical test of the importance of ensuring that witnesses do not know how many images will appear in a sequential lineup. The data suggest that it is absolutely essential when conducting sequential lineups that the witness does not know how many photographs are to be seen. The logic of Lindsay and Wells (1985) was borne out under empirical testing. When participants knew that they were nearing the end of the lineup, they became more willing to choose. Consequently, suspects were more likely to be chosen if they appeared later in the lineup than if they appeared early in the lineup, and this effect was especially strong for innocent suspects. The conditionalized probability of choosing an innocent suspect who was placed last in a nonbackloaded lineup was almost 50%, compared with 10% when placed in Position 2. As a result, the provable value of suspect identifications from nonbackloaded lineups depended on where in the lineup the suspect was placed—clearly an unsatisfactory state of affairs. Fortunately, backloading reduced these position effects, such that suspects were chosen with similar frequency regardless of their position in the lineup. The practical implications of these data are clear: Witnesses must not be aware of the number of images to appear in a sequential lineup. Failing to meet this basic stipulation of the sequential lineup procedure greatly increases the risks to innocent suspects, especially if placed later in the lineup.

The debate over whether the sequential lineup should be advocated to policymakers has intensified over recent years. Critics argue that we have little knowledge of how each element of the procedure adds to its effectiveness because the components have not been subjected to empirical testing (Malpass et al., 2008; McQuiston-Surrett et al., 2006). It is not difficult to imagine how some aspects of the sequential procedure (such as backloading) could become lost in translation from the laboratory to the field, particularly if we have no data to back up our assumptions about the roles that these details play. Arguably, there has never been a more crucial time to scrutinize the various components of the sequential (and the simultaneous) lineup. The Supreme Court of New Jersey recently produced a landmark ruling regarding eyewitness identification procedures, giving defendants greater powers to challenge potentially biased identification procedures. Field tests are currently underway in the United States to test the reliability of the sequential lineup with real witnesses (see the report for the American Judicature Society by Wells et al., 2011), the findings of which could have widespread impact on practice and policy. We must work toward a better understanding of how the sequential—and indeed any other—procedure works to ensure that all suspects face the fairest possible test of their guilt or innocence.

References

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes. Available at www.r-project.org.


Received January 15, 2012
Revision received July 5, 2012
Accepted July 10, 2012