Measuring the Goodness of Lineups: Parameter Estimation, Question Effects, and Limits to the Mock Witness Paradigm

GARY L. WELLS* and AMY L. BRADFIELD

Iowa State University, USA

SUMMARY
Lineups and photospreads can be biased against a criminal suspect and there is a need to measure this bias. The mock witness method has been accepted by eyewitness scientists since the 1970s as the paradigm for generating the data on which various metrics of bias are based. We note the reasons that structural lineup bias can lead to mistaken identification and we discuss the statistical metrics that are proposed to measure the degree of bias. Using mock witness data, we show that estimates of bias can vary as a function of the question asked of the mock witnesses. In addition, we note various lineup biases that are not measured with the mock witness paradigm, such as procedural biases and problems with propitious heterogeneity. Copyright © 1999 John Wiley & Sons, Ltd.

An eyewitness’s identification of a criminal suspect from a lineup or photospread can be a powerful event in terms of convincing triers of fact that the suspect was the perpetrator of the witnessed crime. Experimental studies, for example, show that people are readily persuaded by eyewitness identification testimony, especially if the eyewitness is perceived as confident (e.g. Cutler et al., 1988; Fox and Walters, 1986; Wells et al., 1979, 1981). Studies of actual cases of wrongful convictions by juries show that eyewitness identification testimony was the common factor in over 50 per cent of these convictions (e.g. Connors et al., 1996; Huff et al., 1986; Wells et al., 1998). Mistaken identifications can arise from two sources: (1) variables beyond the control of the criminal justice system, such as the natural limitations of human perception and memory or poor witnessing conditions, (2) variables that are under the control of the criminal justice system, such as the procedure used to obtain the identification from the eyewitness (Wells, 1978). Our concern is with the latter, which have been termed ‘system variables’. Among the best known of the system variables is lineup bias, which we define as characteristics of the lineup that make an innocent suspect stand out in ways that could increase the chances that the suspect is
identified. The classic example of lineup bias is when the suspect is the only person in the lineup who fits the description that the eyewitness had given of the perpetrator. For example, when an eyewitness describes the perpetrator as being a tall, moustached, young male and the suspect is the only one who fits that description (the others being short, or without moustache, or middle-aged), then the structure of the lineup is said to be biased against the suspect.

WHY IS STRUCTURAL LINEUP BIAS A PROBLEM?

Why should an innocent suspect be at risk of mistaken identification merely because he or she is the only one who fits the eyewitness’s pre-lineup description of the perpetrator? Three related, but conceptually different, processes have been posited to explain why structural bias places an innocent suspect at risk.

Inference processes

Police investigators presumably have chosen the suspect, in part, because he or she fits the description that the eyewitness had given of the perpetrator. Given this assumption, it is easy to see how the eyewitness could conclude that the person who fits the description is the one who the police believe is the perpetrator. Hence, rather than using recognition memory to identify the suspect, the eyewitness could simply infer the hypothesis of the police investigators (‘It must be number four’) and use that inference to make the choice. It is useful in this regard to note how conducting a lineup is analogous to conducting a psychological experiment (see Wells and Luus, 1990, for a more complete articulation of this analogy). Precautions that are taken in a psychological experiment should also be taken in conducting a lineup. In a psychological experiment, for example, we go to great lengths to make sure that the experimental hypothesis is kept from the participants. This way, we know that the participants’ responses do not reflect merely what they think the experimenter believes. One situation in which the police hypothesis becomes obvious to the eyewitness is when the suspect is the only lineup member who fits the eyewitness’s previous verbal description of the perpetrator.

Relative judgements

A second, related, process that comes into play if the suspect is the only one who fits the description concerns the concept of relative judgement processes. It has been shown that eyewitnesses tend to select the person who most closely resembles their memory of the perpetrator relative to the other lineup members (see Wells, 1984, 1993). Wells (1993), for instance, examined the rate at which eyewitnesses to a staged crime selected the perpetrator and the next-most-popular choice from a perpetrator-present six-person lineup. The perpetrator was chosen by 54 per cent and the next most common selection was chosen by 13 per cent. When the perpetrator was removed, however, the person who was chosen by 13 per cent when the perpetrator

2This type of bias, which we call structural bias because it is a characteristic of how the lineup is structured, is not the only type of system-variable bias that can inflate the chances that an innocent suspect is identified. As we note later, there are also procedural biases that can have similar effects.
was present was chosen by 38 per cent when the perpetrator was not present. Even though the witnesses were warned that the actual perpetrator might not be in the lineup, they tended to simply identify the next best choice when the actual perpetrator was removed. This result vividly demonstrates relative judgements at work; eyewitnesses tend to select the person who most resembles their memory of the perpetrator relative to the alternative people in the lineup. If the suspect is the only one who fits the description, eyewitnesses will tend to identify that person merely because he is the best choice among the choices offered. When all members of the lineup fit the general description that the eyewitness had given of the perpetrator, on the other hand, it would take a coincidence (of chance $1/N$, where $N$ is the number of lineup members) for an innocent suspect to be a better fit to the eyewitness’s memory than the other lineup members.

The uncertainty-reduction function of a lineup

A final, theoretical, reason why it is a problem if the suspect is the only one who fits the verbal description that the eyewitness gave of the perpetrator concerns the general purpose or intended function of eyewitness identification tasks. As Luus and Wells (1991) noted, the purpose of conducting a lineup is to learn something more than what was already known from the eyewitness’ verbal description of the perpetrator. In other words, a lineup should further reduce uncertainty about the guilt of the suspect beyond what we already know from the verbal description. The implicit theory behind lineups is that the eyewitness has information in recognition memory (tapped by the identification task) that goes beyond the level that was obtained from the recall task (description) and that this additional information can further reduce uncertainty as to the identity of the perpetrator. Accordingly, the recognition task needs to be structured so that it is not simply redundant with the recall task. This means that the recognition task needs to control for recall by ensuring that all lineup members fit the eyewitness’s recall (verbal description) equally well.

In summary, there are three theoretical reasons offered as to why it is problematic if the suspect is the only person who fits the description: (1) it allows the eyewitness to make inferences regarding which person the police think is the suspect, (2) it guarantees that the suspect, guilty or not, is the relative best fit to the witness’ memory, and (3) it fails to go beyond mere recall memory and does not require the witness to rely on recognition memory.

**EMPIRICAL EVIDENCE THAT STRUCTURAL LINEUP BIAS IS A PROBLEM**

Experimental studies have clearly shown that an innocent suspect is at risk of being falsely identified when the suspect is the only one who fits the eyewitness’ description of the perpetrator. In their staged-crime studies, Lindsay and Wells (1980) found that the innocent suspect was nearly two and one-half times more likely to be falsely identified when he was the only person in a lineup who fit the description than when there were at least two other people who fit the description. Using a staged-crime methodology in which the participant-eyewitnesses believed that they were involved in a real case, Wells *et al.* (1993) found that false identifications rose from 10 per cent
when all six lineup members fit the description to 45 per cent when the suspect was the only one who fit the description. (Note: The chances of mistaken identification in culprit-absent lineups is not equivalent to \(1/N\) where \(N\) is the number of lineup members who fit the description because none of the above responses are explicitly allowed.)

It appears that another effect of structural lineup bias could be the inflation of confidence. In the Wells et al. (1993) study, the mean confidence for eyewitnesses who made identification attempts from the biased lineup was 4.5 on a 7-point scale, significantly higher than the mean confidence of 3.7 for those who attempted their identifications from the unbiased lineup. This means that the type of lineup that produces the greatest chance of false identification also produces the greatest level of confidence. Although not predicted at the time, this finding makes sense from our theoretical understanding of why biased lineups are a problem. As indicated previously, the biased lineup conveys information to the witness about which person is the suspect, leads the suspect to stand out as the closest fit to the witness’s memory, and does not require the witness to perform a true recognition task. In short, the task of making an identification from a biased lineup probably appears to be an easy one, thereby leading the eyewitnesses to be more confident in their decision even while being more likely to make an error.

Importantly, these same studies have shown that the chances that the suspect will be identified if he or she is the actual perpetrator (i.e. ‘hit’ rates) are not appreciably harmed by making the lineup unbiased. For example, in the Wells et al. (1993) study the rate of accurate identifications of the perpetrator in perpetrator-present lineups was 71 per cent with the biased lineup and 67 per cent with the unbiased lineup. This is an important observation because it means that unbiased lineups can actually increase the diagnostic value of the lineup by reducing false identification rates while holding accurate identification rates constant. This net benefit in diagnostic value from unbiased lineups is probably due to the way that unbiased lineups force the eyewitness to use recognition memory rather than inference processes.

MEASURING STRUCTURAL LINEUP BIAS

Because lineup bias is known to affect the chances of a mistaken identification, a premium has been placed on measuring this bias. If we can measure it, we can provide assessments of the presence and absence of such a bias in real cases. In one sense, the presence of lineup bias in an actual case constitutes an alternative interpretation for the eyewitness’s identification of a suspect (i.e. an alternative to the interpretation that the suspect is the perpetrator). Simply stated, rather than recognizing the suspect as the perpetrator, the eyewitness might have selected the suspect because he or she ‘stood out’, which is an inference process rather than a memory process.

How can lineup bias be measured in actual criminal cases? In actual cases, there is virtually always a picture taken of the lineup (or, in the case of photospread identifications, a copy of the photospread) to examine. Doob and Kirshenbaum (1973) proposed a ‘mock witness’ paradigm to measure lineup bias. Unlike the actual witness, mock witnesses have never seen the perpetrator, but their task nevertheless is to view a photo of the lineup and select the person they think is the suspect. Before making a selection, the mock witnesses are provided with the pre-lineup verbal
description that the witness had given of the perpetrator. The number of mock witnesses who choose the suspect versus the number who select other lineup members constitutes data from which inferences can be made regarding the presence of lineup bias.

Using mock witness data, Wells et al. (1979) proposed a measure (or metric) called functional size. Functional size simply is the ratio of total mock witnesses making a lineup selection to the number who select the suspect. Hence, if 100 mock witnesses view a lineup and 20 select the suspect, functional size is 5.0; if 25 select the suspect, then functional size is 4.0; if 50 select the suspect, then functional size is 2.0, and so on. The nominal size of the lineup (the number of persons in the lineup regardless of whether or not they were selected by mock witnesses) is not pertinent to the calculation of functional size. Although functional size can be compared to nominal size, functional size rather than nominal size is presumed to be a measure of lineup bias. For example, two lineups, one having nominally ten members and the other nominally five, are equivalent in bias if they are equivalent in functional size. Along these lines, two lineups that have the same nominal size (e.g. ten persons) are not equally biased if they have different functional sizes. Risk of mistaken identification owing to lineup bias is then presumed to be captured by functional size independently of the nominal size of the lineup.

Other metrics have been proposed as potentially superior estimates of lineup bias, in particular Malpass’s measure called defendant bias (see Malpass, 1981; Malpass and Devine, 1983). This metric is somewhat more complex than functional size because the frequency of selections of each lineup member must be compared to calculations of chance expectations. We use the functional size metric in our examples in this article because of the ease of calculation and because it has become a common metric in the eyewitness literature for measuring lineup bias.

In using functional size, we do not mean to suggest that it is superior to defendant bias as a measure of bias. In fact, we believe that perhaps the best metric of lineup bias is simply the proportion of mock witnesses who select the suspect. We note, for example, that both functional size and defendant bias are non-linear transformations of the proportion of mock witnesses who select the suspect. We see no particular reason to think that any non-linear transformation of the proportion would be a better index than the proportion itself. Hence, for the data reported in this article, we report not only functional size but also the proportion of mock witnesses who selected the suspect.

Ultimately, the question of what metric should be preferred as a measure of lineup bias should be determined by the answer to the question of which metric best predicts the identification behaviors of eyewitnesses. This is a problem that Rod Lindsay and his colleagues have tackled recently (see Lindsay et al., this issue). Their work shows that the magnitude of the correlation between the identification behaviours of eyewitnesses and the proportion of mock witnesses selecting the suspect is greater than the correlation between the identification behaviours of eyewitnesses and any other metric of bias, including functional size.

**COURTROOM PRESENTATION ISSUES**

The purpose of this article is to focus on the mock witness data collection method itself, rather than the statistical properties of the metrics one uses to express the data.
Nevertheless, there are a few points about lineup bias that need to be emphasized as they relate to matters of courtroom presentation. First, lineup bias is not a property of the lineup *per se*. Instead, bias is an interaction between the lineup and the verbal description that the witness had given of the perpetrator. The same lineup could yield different proportions of mock witnesses selecting the suspect depending on the description that was given (see Corey *et al*., this issue). For this reason, the same lineup could have different functional sizes for different witnesses in a multiple-witness case. Accordingly, it could prove to be rather confusing to a jury if they were told that the same lineup yielded a functional size of 2.0 for one eyewitness and 3.0 for another eyewitness.

Although some circumstances could be relatively simple for the juror, such as when there is only one eyewitness and the functional size of the lineup is 1.0, real cases have a way of being rather messy. For instance, it is not uncommon for there to be one eyewitness who identified the suspect and a second eyewitness who saw the same lineup but made no identification or identified a distractor from the lineup. Assume that the functional size of the lineup for both eyewitnesses was 1.0 and an eyewitness expert now tries to make an argument that the lineup was biased. To what extent does the second eyewitness’s behaviour call into question the argument that there was significant bias in the lineup? If there was such bias, why did the second eyewitness not also select the suspect? We simply do not know how jurors reason about these matters. Through experience, we have learned that real cases generally are much more complex than our hypothetical examples and that we cannot test all possible situations in terms of how jurors will reason about such matters. What is needed is a general theory of how jurors reason about eyewitness issues rather than a haphazard collection of specific findings.

Another note about presenting lineup bias information in court is that, like all measures, the data are at best an approximation to the true value. Among other things, this means that there is error in the obtained value and this information must be incorporated into the testimony given to a jury. Although there can be systematic error, our concern at this point is with random error. A primary factor governing random error in the mock witness paradigm is sample size.

We have been concerned about sample size recently because we have noted that some eyewitness experts have presented estimates of lineup bias based on sample sizes far below what would be needed to make a confident claim. In the examples that follow, we show the effect of sample size on error in estimates of functional size. However, we could show these same sample-size effects for other measures, such as the proportion measure, defendant bias, and acceptable foils. Consider, for example, an expert who collects mock witness data from 30 people and then concludes that the functional size of a lineup is 2.0. This conclusion ignores the concept of random error. Any given point estimate from sample data has a calculable 95 per cent confidence interval. The 95 per cent confidence interval around a functional size of 2.0 based on a sample size of 30 mock witnesses includes the values 3.0 and 1.0. Hence, we suggest that sample sizes be sufficient to bring the confidence interval to $\pm 0.5$ on functional size.

It is important to realize that the sample size required to bring the 95 per cent confidence interval to $\pm 0.5$ on functional size is not a constant number. Instead, the sample size required varies as a function of the obtained values. Furthermore, the confidence interval is not symmetric around the obtained value. Suppose, for
example, 50 mock witnesses were used and 25 chose the suspect, yielding a functional size estimate of 2.0. In this case, the 95 per cent confidence interval is 1.6–2.8. Hence, we can be 95 per cent confident that the value is closer to 2.0 than to 1.0, but we cannot be 95 per cent confident that the value is closer to 2.0 than it is to 3.0 because the interval includes the value 2.5. When the obtained value is 2.0, it takes about 80 mock witnesses to conclude with 95 per cent confidence that the functional size is closer to 2.0 than it is to 3.0. Consider now an obtained functional size estimate of 4.0. In this case, 50 mock witnesses yields a 95 per cent confidence interval of 2.7–7.8. Hence, one cannot conclude with 95 per cent confidence that the functional size is closer to the obtained value of 4.0 than it is to 3.0 or 7.0. That would be an unacceptable level of precision. In the case of an obtained functional size of 4.0, it takes about 175 mock witnesses to conclude with 95 per cent confidence that the functional size is closer to 4.0 than it is to 5.0. To the extent that an expert is going to assert in court that the functional size of a lineup is equal to some particular value, the expert needs to first calculate the error rate so as to not mislead the triers of fact regarding the precision of the estimate.

These issues raise some important questions regarding what a jury can and cannot understand with regard to expert testimony on assessments of lineup fairness. In some respects, eyewitness scientists’ understanding of lineup fairness is much more advanced than is our understanding of jurors’ abilities to follow the logic of expert testimony regarding lineup fairness.

THE MOCK WITNESS PARADIGM: WHAT IS THE APPROPRIATE QUESTION?

We believe that the mock witness paradigm in its general form is a meaningful and important element in testing the fairness of a lineup. The logic behind the paradigm is unquestionably sound. If a lineup has no biases against the suspect, then people who have never seen the culprit (i.e. are not actual witnesses) should not be able to select the suspect from the lineup when they are armed only with the description that the eyewitness had given of the perpetrator. Of course, some mock witnesses could pick the suspect by chance. Hence, the concern is with the number of mock witnesses who choose the suspect relative to those who select someone else. Most previous work with the mock witness paradigm has concerned itself with the question of what to do with the data after they have been collected, i.e. what is the appropriate metric of fairness? We introduce another concern, however, namely what is the appropriate question to ask of the mock witness?

It appears that researchers have ignored the possibility that the question asked of mock witnesses could affect the results of a mock witness test. Indeed, many researchers have reported functional size, acceptable foils, defendant bias and other measures using the mock witness paradigm without even mentioning the precise form of the question asked of the mock witnesses. In the earliest uses of the mock witness paradigm, mock witnesses were given the description, shown a copy of the lineup or photospread, and asked to pick the person they thought to be the accused (Doob and Kirshenbaum, 1973; Wells et al., 1979). There are other possible questions, however, such as asking the mock witnesses to pick the person who best fits the description or...
asking them to pick the person that they think the eyewitness identified. Are these equivalent questions? We conducted a simple study to find out.

**METHOD**

Five photographic reproductions of live lineups and five copies of photospreads that were used in actual eyewitness cases were shown to 210 undergraduates who served as mock witnesses in the study. All 210 mock witnesses were shown all five lineups and all five photospreads in one of three random orders. Orders were equally represented in all three versions of the question manipulation. Order had no effect on the results and will not be discussed further. Each lineup or photospread had a verbal description of the perpetrator that was given by an actual eyewitness in the case. The description was attached to the bottom of each lineup. All mock witnesses were informed that an eyewitness to a crime had described the perpetrator using the description at the bottom of each photo. One third of the mock witnesses were asked ‘Which person in the lineup do you think is the accused?’, one third were asked ‘Which person in the lineup do you think the eyewitness was describing?’ and one third were asked ‘Which person in the lineup best fits the description?’ We used the forced-choice version of the mock witness procedure (i.e. a none-of-the above option was not provided). Each mock witness answered the same question for each of the ten photos.

**RESULTS AND DISCUSSION**

Table 1 gives the frequencies (out of a possible 70) with which the suspect was selected as a function of the question asked for each of the ten lineups. We conducted chi square tests for each lineup and found that the form of the question affected rates of choosing the suspect for five of the ten, each \( \chi^2 (2 \text{ df}, \ N = 70) > 5.99, \ p < 0.05 \). The five lineups that showed significant effects for type of question are in bold type in Table 1. The general pattern for these significant question effects is that the question ‘Which person best fits the description?’ tends to yield a functional size that is larger than the other two questions. Across all ten lineups, the mean functional size estimate was 4.6 for the description-fit question, 2.9 for the question ‘Which person was the witness describing?’ and 2.5 for the question ‘Which person is the accused?’

What accounts for the effect of question? In what sense should the question ‘Which person best fits the description?’ be different from the other two questions? Furthermore, given that there are differences, which question should be used with the mock

---

3This raises another interesting question as to whether or not it matters if mock witnesses are allowed to indicate ‘none of the above’. In the work of Wells et al. (1979), which originally proposed the functional size metric, mock witnesses were allowed to indicate ‘none of the above’ and 32 per cent chose that option. Functional size estimates were based on the 68 per cent of mock witnesses who made a selection. It is possible that a forced-choice mock witness would yield the same estimate of functional size as would a procedure that allows mock witnesses to make no choice. This could easily be tested by using the procedure in which eyewitnesses are allowed to make no choice and then forcing those who made no choice to make a choice to see if their pattern of choosing differs from those who made a choice. If there are no differences, the forced-choice procedure would clearly be preferred because of the efficiency of having data from every mock witness.
witness paradigm? In some ways we can only speculate why the question effect occurs because we did not take process measures (e.g. think-aloud tasks) that would allow us to fully discover the different strategies of selection that these questions evoke. However, we believe that at least one difference can be understood by considering the concept of asymmetric judgements of similarity. There are some judgement tasks for which an object \(X\) is judged to be more similar to an object \(Y\) than object \(Y\) is to object \(X\). Asymmetries of this sort tend to happen when the salient qualities of \(X\) tend to fit with qualities of \(Y\) better than the salient qualities of \(Y\) fit with \(X\). For instance, a salient associate of 'Californian' is someone who lives in a warm weather climate. Hence, when asked how similar Californians are to Texans, people can see similarities because of the warm weather. However, warm weather is not the most salient associate of 'Texan'. Instead, 'cowboy' or similar associates come to mind when thinking of Texans. Hence, when asked how similar Texans are to Californians, people can see less similarity than when asked how similar Californians are to Texans.

The question effect in the mock witness task can perhaps be understood by this notion of judgement asymmetries. A description might fit a person’s face better than a person’s face fits the description. To illustrate this point, we have reproduced lineup number 7 in Figure 1. The description given by the eyewitness was ‘Black, male, mid-20s, facial hair’. Given this description and asked which person best fits the description, mock witnesses preferred lineup member number three (on the far right), presumably because his facial hair was closer to the idea of a beard than is number two. However, when mock witnesses were asked which person the eyewitness was describing, they shifted their preference to lineup member number two. Why would they shift away from number three when asked this question? We think it is because the mock witnesses asked themselves ‘How would I have described number three if I had seen him?’ and decided that they would have described his most unique feature of having curls in his hair. Given that the eyewitness said nothing about the unique hair, it seemed likely that number three was not the person that the eyewitness was describing. Hence, although number three seems to fit the description better than

<table>
<thead>
<tr>
<th>Lineup number (and nominal size)</th>
<th>Which is accused? Number choosing suspect (and functional size)</th>
<th>Which was witness describing? Number choosing suspect (and functional size)</th>
<th>Which best fits description? Number choosing suspect (and functional size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (7)</td>
<td>45 (1.6)</td>
<td>41 (1.7)</td>
<td>39 (1.8)</td>
</tr>
<tr>
<td>2 (4)</td>
<td>36 (1.9)</td>
<td>39 (1.8)</td>
<td>22 (3.2)</td>
</tr>
<tr>
<td>3 (6)</td>
<td>12 (5.8)</td>
<td>12 (5.8)</td>
<td>6 (11.7)</td>
</tr>
<tr>
<td>4 (6)</td>
<td>45 (1.6)</td>
<td>23 (3.0)</td>
<td>19 (3.7)</td>
</tr>
<tr>
<td>5 (5)</td>
<td>38 (1.8)</td>
<td>33 (2.1)</td>
<td>37 (1.9)</td>
</tr>
<tr>
<td>6 (6)</td>
<td>55 (1.3)</td>
<td>50 (1.4)</td>
<td>52 (2.8)</td>
</tr>
<tr>
<td>7 (3)</td>
<td>49 (1.4)</td>
<td>32 (2.2)</td>
<td>21 (3.3)</td>
</tr>
<tr>
<td>8 (5)</td>
<td>23 (3.0)</td>
<td>25 (2.8)</td>
<td>25 (2.8)</td>
</tr>
<tr>
<td>9 (10)</td>
<td>42 (1.7)</td>
<td>45 (1.6)</td>
<td>20 (3.5)</td>
</tr>
<tr>
<td>10 (11)</td>
<td>15 (4.7)</td>
<td>19 (3.7)</td>
<td>6 (11.7)</td>
</tr>
<tr>
<td>Means</td>
<td>36.0 (2.5)</td>
<td>27.4 (2.9)</td>
<td>24.7 (4.6)</td>
</tr>
</tbody>
</table>

Note. Rows in bold type indicate lineups for which the form of the question produced significant differences in the rates of choosing the suspect.

Measuring the Goodness of Lineups

Figure 1. Three-person lineup number 7, which showed an effect for type of question asked of mock witnesses
number two, the description itself seems to be most likely to have been given for number two rather than for number three.

We think that there could be yet another difference among these three questions. Suppose that the suspect does not stand out from the other lineup members based on the description itself, but stands out for some other reason. Suppose, for instance, that the suspect is the only one facing the left, or the only one who appears to be nervous. The question ‘Which person best fits the description?’ is insensitive to this possible bias whereas the question ‘Which person is the accused?’ might reveal the bias. This appears to be the case for photospread number four, for which the suspect was the only one whose photo had been darkened. Notice that this is the photospread for which the question ‘Which person is the accused?’ yielded a smaller functional size estimate (1.6) than did the question of ‘Which person was the witness describing?’ (3.0).

These data, and our interpretation of the differences, lead us to recommend that the question asked of mock witnesses should be of the form ‘Which person is the accused?’ We believe that this question allows the eyewitness to use not only the description but also other information that was available to the eyewitness (i.e. other things that might have made the suspect stand out).

WHAT THE MOCK WITNESS PARADIGM DOES NOT MEASURE

Continued refinements of the mock witness paradigm, development of the best metrics, and tests of assumptions underlying the paradigm are important to eyewitness scientists’ abilities to assess lineups, understand lineup biases, and direct the justice system on how to eliminate biases. However, it is essential to recognize that there are biases that the mock witness method cannot measure. Consider, for instance, live lineups in which the photograph fails to capture certain behaviours of the lineup members. Usually, the suspect in a lineup knows he is a suspect and the distractors know that they are not suspects. Hence, the behaviours of the distractors might be relaxed and calm while the behaviours of the suspect might be rigid and nervous. A videotape of the lineup might capture these qualities, but a photograph is merely a moment in time and might not capture some of these qualities. Even more problematic is the fact that the mock witness paradigm is not capable of measuring procedural biases.

Procedural biases

Procedural biases can be defined as any behaviours on the part of the person administering the lineup that might increase the propensities of the eyewitness to either select the suspect or to have inflated confidence in the selection. It is not our purpose to review procedural biases here, but procedural biases include failure to instruct eyewitnesses that the actual perpetrator might not be in the lineup (Malpass and Devine, 1983, see meta-analysis by Steblay, 1997), failure to use double-blind procedures in administering lineups (Wells and Luus, 1990), and giving eyewitnesses feedback about their identifications prior to assessing their confidence (Wells and Bradfield, 1998, 1999). These biases can be at least as powerful as structural biases,
but, unlike structural biases that can be captured by photographs, procedural biases are difficult to discover in actual cases.

Propitious heterogeneity

Another property of lineups that is not captured by the mock witness procedure is the degree of similarity among lineup members that exceeds what is necessary for obtaining a given functional size. The idea of propitious heterogeneity refers to the fact that not all heterogeneity (differences) among lineup members is sinister. Some differences among lineup members are beneficial (propitious) because some amount of heterogeneity is necessary in order for human recognition memory to work (Wells, 1993). Imagine, for instance, two lineups: one lineup has six people, each of whom fit the description and none of whom stand out for any other reason; the other lineup is composed of the same suspect embedded among his clones. In both cases the lineup has a functional size of six. But, are they equally good lineups? No. Only in the former lineup can we reasonably expect the eyewitness to perform at a level above chance. The clones hypothetical is extreme, but empirical data show that such extremes need not be reached in order to show that eyewitnesses’ abilities to select the culprit in culprit-present lineups is harmed by too much similarity (see Wells et al., 1993). Most eyewitness researchers have not concerned themselves with this problem because a lineup that lacks propitious heterogeneity is not biased against the suspect. Nevertheless, it is important to note that the fairness of a lineup, as defined by the mock witness method, cannot be equated with a broader concept of the ‘goodness’ of a lineup. The goodness of a lineup requires us to consider not only the extent to which the lineup serves to protect the innocent suspect but also the extent to which it allows the eyewitness to identify the actual culprit when the actual culprit is present.

GENERAL DISCUSSION AND CONCLUSIONS

The idea that one can measure structural biases in lineups is one of the most fundamental and best-known premises in eyewitness identification. The mock witness task has and will continue to be the primary method for analysing structural biases in lineups and photospreads in actual cases. While there are emerging reasons to prefer one metric or another for summarizing the data from mock witnesses (see Lindsay et al.’s article in this issue), we have noted that the paradigm itself needs some additional attention. For instance, we need to know whether patterns of choice by mock witnesses differ as a function of whether the task is forced choice versus one that permits mock witnesses to make a none-of-the-above response. In addition, we have shown that the type of question asked of the mock witnesses can affect the pattern of choices. We recommend that the question ‘Which person is the accused?’ be used with the mock witness paradigm rather than a question about the fit between the description and the lineup members. Finally, we note that the mock witness method cannot measure other important biases, such as procedural bias, and can measure only the bias that is captured by the click of a camera at a moment in time.
REFERENCES


