APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **13**: S9–S26 (1999)

Statistical Considerations when Determining Measures of Lineup Size and Lineup Bias

COLIN TREDOUX*

University of Cape Town, South Africa

SUMMARY

Much eyewitness research has centred on the development and evaluation of measures of lineup properties. These measures have proved useful to researchers and to expert witnesses who testify on eyewitness testimony. However, the inferential statistical properties of the measures themselves are rarely taken into account, despite the fact that most of them are derived from the mock witness task, which relies on an implicit probability model. This is a failing, and it is argued that estimates of lineup properties reported without inferential considerations may be misleading. Methods are suggested here for reasoning inferentially about lineup measures, and for planning studies to ensure adequate statistical power. Suggestions made by eyewitness researchers regarding the practical interpretation of lineup measures are critically evaluated, as are recommendations for new lineup measures, as suggested by Lindsay *et al.* (this issue). Copyright © 1999 John Wiley & Sons, Ltd.

Legal psychologists have contributed a valuable corpus of research around the development and evaluation of lineup measures. Measures proposed thus far are of lineup fairness, lineup size, the evaluation of individual foils, and the diagnosticity of particular lineups. The measures are determined through the use of the mock witness technique, or in the case of diagnosticity, simulated or live lineups. The measures are usually of a descriptive nature, and little has been written about their inferential use. It is important, though, to develop ways of reasoning inferentially about the measures, as the measures are all susceptible to random sampling variation – indeed, in the case of the mock witness method, all measures derive their meaning from a probabilistic conceptualization. If we do not do this, we mask the inherent variability and sampling error around our measures, and we may mislead the people to whom we present these estimates.

THE NEED FOR STATISTICAL INFERENCE WITH LINEUP MEASURES

The earliest and least embellished measure of lineup fairness is the proportion of mock witnesses who identify the suspect, and the comparison of this proportion to

*Correspondence to: Colin Tredoux, Department of Psychology, University of Cape Town, Rondebosch 7701, South Africa. E-mail: plato@psipsy.uct.ac.za

CCC 0888–4080/99/SI00S9–18 $17.50

that expected under an assumption of equiprobabilistic choice. I will discuss the mock witness procedure later in the paper, but for the moment I simply want to consider what situations we might end up in if we determine the proportion measure of fairness without bearing some basis principles of statistical inference in mind.

Table 1 presents an array of frequencies representing choices made by 21 mock witnesses, when faced with a lineup consisting of the suspect and five foils. A glance at the table of raw frequencies shows that almost 40% of the mock witnesses choose the suspect, and that the suspect is chosen at a rate which is almost three times that of any of the foils. In addition, the rate at which the mock witness is chosen is well above that of chance expectation. However, if we calculate a 95% confidence interval around the estimate of the proportion, we find that the interval includes the expected value of 0.17. Indeed, the confidence interval is very wide, and we should be very cautious about how we interpret the point estimate of the proportion.

Table 1. A hypothetical frequency array of mock witness choices, exhibiting the effect of sampling variation

	Foil 1	Suspect	Foil 2	Foil 3	Foil 4	Foil 5
Frequency	2	8	2	3	3	3

Proportion who choose suspect = 0.38
Proportion expected, assuming equiprobabilistic choice = 0.17
95% Confidence Interval on proportion choosing suspect = (0.17; 0.59)
Functional size = 2.61 95% CI = (1.7; 5.88)
Effective size = 4.71 95% CI not calculable

It is relatively easy to see why this is the case: there are only 21 mock witnesses, and there are six lineup members. Under these conditions, the sampling error of the proportion is very high, and we need to show a commensurate amount of caution in how we interpret this proportion. (The numbers of mock witnesses and lineup members in this example are not simply chosen for emphasis – the numbers correspond to those used by Doob and Kirshenbaum, 1973.)

This example suggests that we should be mindful of certain statistical considerations when we determine and report the measure of lineup fairness devised by Doob and Kirshenbaum (1973). It is not enough to say that 38% of a sample of mock witnesses chose the suspect – we need to go further, and report the amount of sampling error we incurred by virtue of the procedure we used. In fact, it may be a bit misleading to simply report the point estimate, in the sense that it can lead to questionable conclusions.

I suggest that it is not only important to do this with the measure of fairness suggested by Doob and Kirshenbaum, but also with other lineup measures, most of which rely on the mock witness procedure. Thus, when we calculate measures of functional size, effective size, and defendant bias, we also need to take statistical considerations into account, because those measures are also subject to random sampling variation.

What I will do in the remainder of this paper is (1) suggest some ways in which we can apply statistical inference to lineup measures, and (2) comment on the contributions made by other authors in this symposium to the measurement of lineup fairness. The suggestions for applying methods of statistical inference will of necessity

Table 2. Suggestions of inferential procedures to apply to lineup measures

Lineup measure	Suggestion
Probability that k of N mock witnesses choose the suspect	Calculate exact p values from Binomial probability distribution
Proportion identifying suspect (Doob and Kirshenbaum, 1973)	Use confidence interval [CI] (normal approximation to the binomial)
Functional size (Wells *et al.*, 1979)	Use CI on proportion, take reciprocals of endpoints of interval
Effective size (Malpass, 1981)	Use an alternative measure, I (Agresti and Agresti, 1978), with a modification
Defendant bias (Malpass, 1981)	As for estimates of proportions
Foil feasibility (Malpass, 1981)	Construct CIs around estimates of foil proportions, use endpoints as feasibility criteria
Diagnosticity (Wells and Lindsay,1980)	Treat diagnosticity as a relative risk ratio, and apply known methods (Rothman, 1986)

be very brief, but a more comprehensive account is provided in Tredoux (1998). The suggestions are summarized for ease of convenience in Table 2.

THE MOCK WITNESS PROCEDURE

Most lineup measures derive from or depend on the mock witness procedure, so it is appropriate to think about ways of applying statistical inference to the procedure. This is, in fact, quite straightforward, since the mock witness task is much like a coin or die-throwing experiment, with a fixed number of trials. The mock witness task requires that subjects blind to the identity of the suspect attempt to identify the suspect from an array of lineup members, using a verbal description.

There are N mock witnesses, and there are k lineup members. The probability that a witness will choose the suspect, given that the choice is made randomly, will be $1/k$. We assume that witnesses choose independently of each other. Then each individual witness choice can be thought of as a Bernoulli trial, with probability of success $= 1/k$. The number of trials, q, in which a successful choice is made, will take the Binomial probability distribution. This distribution can be used to determine the probability that, say, eight of 21 mock witnesses choose the suspect, assuming that they are choosing randomly. Similarly, the cumulative binomial probability distribution can be used to calculate the probability that one witness, or two witnesses, or three, ..., or t identified the suspect just by chance.

There are other methods of evaluating the number of accurate identifications made by mock witnesses. Malpass and Devine (1983), and Buckhout *et al.* (1988), for example, use a one-sample z-test to compare the observed proportion of correct witness choices to that expected under an equiprobability model. It is a simple enough task, but it relies on the fact that the normal distribution is the limiting distribution of the binomial: the approximation is good for large samples, but may not be very good for small ones, especially when the parameter $1/k$ is small. In mock witness experiments one, or both, of these conditions may not be true. I suggest that, in most cases, it is better to use the binomial distribution to evaluate the rate at which mock witnesses identify the suspect. It yields an exact probability estimate, and the underlying probability model accurately represents the nature of the mock witness task.

A frequently expressed criticism of this 'test of lineup fairness' is that the outcome of the test depends directly on the number of mock witnesses. Wells, Leippe & Ostrom (1979), for example, argue that in order to conclude that a lineup is biased, researchers would simply need to obtain a sufficiently large sample and conduct a mock witness evaluation of the lineup. Consequently, we need to view claims of bias supported by mock witness tasks with extreme caution. Brigham, Meissner and Wasserman (this issue) make a similar claim.

The argument is correct, but it applies to almost all instances of significance testing, and not merely to significance test evaluation of lineups. The appropriate use of the mock witness task is dependent on the good judgement of the researchers who apply it, and this is most certainly true of any inferential statistical procedure. Despite the argument made by Wells *et al.* and Brigham *et al.*, a significance test on the proportion of mock witnesses choosing the suspect – or better, perhaps, a confidence interval – provides useful information about a lineup. This is because it takes random sampling variation into account, and gives us an indication of the statistical reliability of the index of lineup bias. If we use 1000 mock witnesses, we find that the suspect is chosen at a rate of 0.19 from a six-person lineup, and test this proportion against that expected by chance, we will correctly conclude that the lineup is biased against the suspect. What the significance test does not, and cannot, tell us is whether this amount of bias is sufficiently large to be practically significant. This decision of 'practical significance' is one that will have to be taken by the fact finders, or the lawmakers, as both Wells *et al.* (1979) and Malpass (1981) have pointed out.

In general, the question of how to interpret lineup measures is somewhat vexed, and I will return to it later in the paper, apropos of the considerations raised by Brigham *et al.* in this issue.

Proportion identifying suspect

I started this motivation for the use of statistical inference on lineup measures by discussing the estimation of proportions, so there is no need to repeat it here, except to note that there are several ways to construct confidence intervals around proportions. One way is to use the normal approximation to the binomial, but this approximation is better when the sample of mock witnesses is large, and it is prudent to consider alternatives (see Hays, 1994).

Functional size

An important feature of a lineup appears to be the number of plausible foils that it contains. Wells *et al.* (1979) coined the term 'functional size' to deal with this type of situation. Functional size is intended to provide an index of the number of plausible lineup members, and is therefore also a measure of lineup fairness. The measure relies on the mock witness task introduced by Doob and Kirshenbaum, but avoids certain problems. It is defined as (N/D), which is the number of functional members of a lineup, hence the term 'functional size'. It is this index that they suggest as a measure of lineup fairness. When functional size and nominal size are identical, the lineup is clearly fair, but this should not be taken as a necessary or sufficient condition of fairness.

Functional size is just a transformation of the proportion of identifications of the suspect, and the same statistical considerations apply to it as to proportions. In particular, the most appropriate way to apply inferential reasoning here is perhaps to express the observed proportion of accurate identifications as a confidence interval, and then to take reciprocals of the endpoints in order to express the interval in terms of so-called functional size.[1]

Effective size

Malpass (1981; Malpass and Devine, 1983), argues for a distinction between *lineup size* and *lineup bias*. Lineup size refers to the number of plausible members that the lineup contains, and it contributes directly to the fairness of the lineup by decreasing the probability that the defendant is identified by a witness who wilfully chooses at random (however unlikely this is). Lineup bias, on the other hand, is the extent to which mock witnesses choose the defendant at rates greater (or smaller) than chance expectation. Because both these components contribute to the fairness of a lineup, a measure of each is required.

The first measure Malpass proposes is 'effective size', which is close in meaning to the measure of functional size. In order to evaluate a lineup, Malpass argues, the critical thing to know is how many plausible foils it contains. The intent of the measure of effective size is to reduce the size of the lineup from a (corrected) nominal starting value by the degree to which members are, in sum, chosen below the level of chance expectation.

The idea behind the measure of effective size is interesting, and importantly, it attempts to use information regarding the distribution of identifications across the lineup, which is necessary if we are to say something about the overall constitution of the lineup. However, there are two features which seem rather problematic.

The first is the assumption that null foils are totally implausible – indeed, non-members of the lineup. This is not the case. Null foils have a positive probability of occurring, especially when the number of lineup members is large and the sample of mock witnesses relatively small.[2] It is better not to reduce the starting number of foils by removing null foils.

The second is that the way the measure is formulated makes the application of statistical inference a little difficult. It appears to be intractable in this respect. However, it is possible to use an alternate measure whose properties are well understood. What I have in mind is a measure of qualitative variation, I, discussed by Agresti and Agresti (1978). I reflects the extent of departure of an array of frequencies from equiprobabilistic expectation, and it is possible to construct confidence intervals around I, as well as to test for differences between independent I's. In addition, it appears to give estimates that are very similar to those produced by effective size, on the same data. The measure is discussed in detail by Agresti and Agresti and by Tredoux (1998).

[1] Rich Gonzalez has reviewed some of the work reported in this paper, and disagrees on this suggestion (November, 1995). He thinks that this will lead to biased estimates of the interval, since the transformation is non-linear, but concedes that the approximation will be better with large samples.

[2] Using the data of Table 1 for example, it can be shown that if witnesses choose randomly the probability that there will be at least one null foil is 0.31.

Defendant bias

Malpass also suggests a measure of lineup bias, which he calls 'defendant bias'. This is similar in concept to the measure proposed by Doob and Kirshenbaum, i.e. the proportion of mock witnesses choosing the suspect is compared to that expected by chance. Defendant bias is different, though, in its definition of chance expectation – instead of defining chance expectation as [1/number of foils], it is defined as [1/effective size]. The assumption is that there are fewer plausible foils than there are nominal foils, and that a better estimate of the likelihood of a mock witness guessing the identity of the suspect should take this into account.

Statistical considerations apply in two respects to this measure. In the first instance, the proportion of mock witnesses choosing the suspect can be expected to exhibit random sampling variation, and in line with suggestions earlier in the paper, could be expressed as a confidence interval. Second, the estimate of effective size used in the calculation of defendant bias can also be expected to exhibit random sampling variation, and it could also be expressed as a confidence interval, under the considerations outlined above. When comparing the confidence interval around the proportion of mock witnesses choosing the suspect to the confidence interval expressing random guessing, situations where there is overlap between the confidence intervals should lead to a conclusion of no bias, and conversely for situations of no overlap, to a conclusion of bias.

Judging the adequacy of individual foils

Malpass and Devine (1983) also suggest a method for evaluating the suitability of individual foils, which is closely related in principle to the measure of effective size. The critical datum for evaluating the suitability of an individual foil, Malpass and Devine suggest, is the extent to which the foil is chosen below chance expectation in a mock witness task. (The question of the extent of departure from chance expectation is a thorny one. Malpass and Devine suggest leaving the decision to the fact finders.) An alternative approach to measuring parade fairness would then be to set a minimum size criterion, and to determine whether the parade meets the minimum size (the estimate of size would be determined by including only plausible foils – and the suspect – in the total).

I suggest here that decisions about whether an individual foil meets some minimum identification criterion should be made in terms of the binomial probability model underlying the nature of the mock witness task, outlined above. This is an important consideration, since we can reasonably anticipate considerable departures from expected rates of identification, especially when relatively small samples are used. For example, imagine that we use a lineup with 10 members, a sample of 30 mock witnesses, and a minimum criterion size of 67%. Then we can apply the cumulative binomial distribution to determine the probability that any particular foil is chosen at a rate below 0.67 of chance expectation, and this value is 0.41. This is a very high probability, and only a large increase in sample size will reduce it to an acceptably low level.

It is therefore perhaps inappropriate to simply disregard foils chosen at rates below some minimum criterion. A better method may be to construct confidence intervals around the observed proportion of identifications that each foil receives, and to apply the minimum criterion test to the endpoint(s) of the intervals. This would have the benefit of attaching some level of probabilistic confidence to any decisions taken about the plausibility of foils. Since intervals would need to be constructed for each foil, it is worth considering some correction for a potential increase in the Type I error rate.

The end of 'lineup size'?

Rod Lindsay and his associates make a concerted argument in this issue against the notion of 'lineup size', especially as operationalized by Malpass (1981) in the measures of effective size and 'number of acceptable lineup members'. Their evidence-in-chief is the failure of the measures to predict the probability of mistaken suspect identification in a set of mock and simulated witness experiments. They conclude that 'There would appear to be little reason to use measures of lineup size given their failure to postdict identification decisions' (p. 10 of the draft copy). This conclusion may be premature. The failure of effective size to predict mistaken suspect identification probability may simply be an artifact of the research design employed by Lindsay *et al.* In order to demonstrate this, it is necessary to carefully review an aspect of the mock witness procedures used by Lindsay *et al.*

The critical lineups for Lindsay *et al.* were those where the target or perpetrator was not present. This allowed them to relate the performance of mock witnesses to mistaken identifications made by simulated witnesses on the same lineup. However, in simulated target-absent lineups there is no suspect, unlike police target-absent lineups where there is (almost) always a particular suspect. In previous studies, eyewitness researchers have somewhat arbitrarily designated one of the lineup members as the suspect, and calculated lineup measures in relation to this elected suspect. Lindsay *et al.* decided instead to treat each of the lineup members as an elected suspect, and calculated lineup measures and suspect identification probabilities for each lineup member. They reasoned that this was justifiable, in the following terms:

> Given that all members of a criminal-absent lineup match the general description, any one of them could be the suspect. Also, real suspects are arrested for reasons other than their match to the description ... These considerations led us to treat each member of the criminal-absent lineups as suspects ... such that each criminal-absent lineup generated six data points ... This permitted us to calculate lineup fairness measures and identification rates for a total of 108 targets ... despite having used only 18 such lineups.

The problem with this procedure, justifiable though it may sound from one line of reasoning, is that data points become non-independent, with profound consequences for some lineup measures, and less profound but significant consequences for other lineup measures, and for attempts to relate these to suspect identification probabilities.

In the first instance, suspect identification probabilities calculated for different members of the same lineup are highly dependent. The sum of these probabilities is unity,[3] and it follows that if a lineup member is identified with high probability, the

[3] This depends on whether the no-choice option is used in the simulated lineup. If it is not used, then the sum is unity, but if it is not, then the sum is $(1 - p(\text{no choice}))$.

other lineup members must be identified with lower probability.[4] This dependence will bias measures of the relationship between suspect identification probability and other indices. This is because identification probability will be constrained to vary by the dependency condition, and will therefore not exhibit the variation it might when tested with independent lineups.

A more severe problem introduced by the mock witness procedure used by Lindsay *et al.*, though, is the effect it has on measures of effective size and 'number of acceptable lineup members'. One of the most important features of the effective size estimator from a statistical point of view is the fact that it takes into account the rate at which witnesses choose foils as well as the rate at which witnesses choose the suspect. In this sense it is a sufficient estimator. It follows that a particular lineup as used in a particular mock witness experiment can generate only one estimate of effective size. In Lindsay *et al.*'s procedure, though, each lineup member is treated as a suspect, and effective size is calculated for each lineup member. The values of effective size calculated in this manner are, as expected, identical. If we understand this problem in a different way, namely in terms of statistical dependence, then we can say that by using the same mock witness procedure for multiple determinations of lineup measures, we create estimates that are statistically dependent. In the case of effective size the estimates are not merely dependent, but identical, since the estimator uses all the available information in the array of frequencies. The consequence of this is that the estimates of effective size calculated by Lindsay *et al.* show a severe restriction of range, and a concomitant reduction in the strength of the relationship between effective size and probability of mistaken suspect identification.

Lindsay *et al.* recognize that estimates of effective size based on the same lineup will be identical, but suggest that this is a problem with the estimator, and one reason for dispreferring effective size as a measure. However, it may simply be a consequence of the particular mock witness procedure used by Lindsay *et al.*, since the procedure renders estimates of lineup size (and lineup bias) statistically dependent. The relationship between lineup size and probability of mistaken identification may turn out to be very different when the data-collection procedure ensures statistically independent estimates of lineup size (and identification probability).

Is it in fact possible – say, for a particular set of mock and simulated witness pairings – that lineup size might bear a strong relationship to probability of mistaken identification, but that this relationship goes undetected when the estimates of effective size and identification probability are collected in the manner reported by Lindsay *et al.*? In order to explore this question, consider the hypothetical set of results displayed in Figure 1, as collected from six mock witness procedures and six simulated witness-identification experiments, where the relevant estimates are determined following the procedure adopted by Lindsay *et al.*

There are six people in each lineup. One of the members in each lineup is arbitrarily identified as the suspect (number 3), and the rest as innocent foils. These lineups are shown to both mock witnesses and simulated witnesses, and effective sizes and identification probabilities[5] are determined, under the procedure used by Lindsay

[4] More accurately, if p(lineup member x) $= \delta$, then p(any other lineup member) $\leqslant (1 - \delta)$.

[5] We will assume for the moment that all simulated witnesses choose a member of the lineup, and none choose the 'not-present' option.

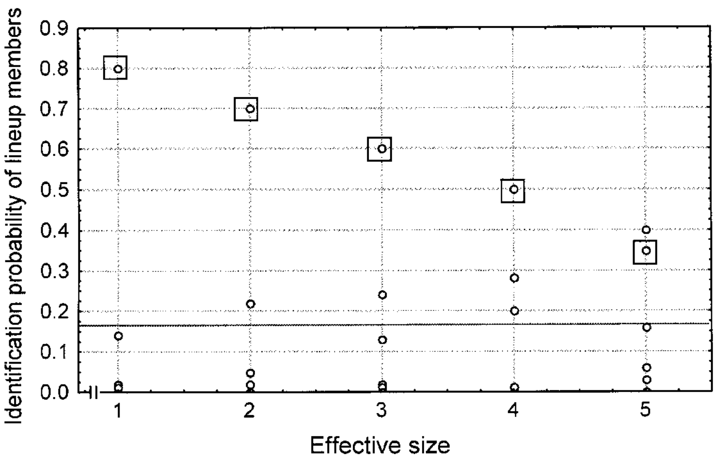

Figure 1. The relationship between effective size and suspect identification probability in a hypothetical set of target-present lineups. Note: The straight line is the least squares regression fit to the individual data points. The data points enclosed in squares are the rates at which suspects are chosen

et al.[6] The figure shows the relationship between effective size and identification probability, and makes several things clear.[7] First, an attempt to fit a least-squares regression line in the *xy* plane shows that there is no relationship between effective size and identification probability. However, the data points with square borders are the frequencies with which the suspect is chosen. It is clear that if these data points are considered on their own, then effective size and mistaken suspect identification probability co-vary strongly. Second, the dispersion of data points within each level of effective size is interesting, because it shows the effect of statistical dependency. The dispersion is much greater when at least one of the lineup members attracts a large number of mock witness choices, since the remaining lineup members can by definition attract only a small number of mock witness choices.

The figure is a clear demonstration that it is possible for effective size to co-vary almost perfectly with the probability of mistaken suspect identification when the estimates are calculated from independent mock witness experiments, and yet fail to co-vary when the estimates are calculated from sets of dependent mock witness experiments. In fact, the figure demonstrates that there are situations – if we use the Lindsay *et al.* procedure – where it is not possible to show co-variation between effective size and mistaken suspect identifications, even when this obviously exists.

To take the question one step further, one can ask whether it is ever possible for lineup size to bear any relationship to probability of mistaken identification if the estimates of effective size and identification probability are collected in the manner reported by Lindsay *et al.* Two pieces of evidence suggest that it might simply not be possible at all. Both are easily conceptualized in relation to Figure 1.

[6] Lindsay *et al.*, however, did not treat any of the lineup members as a fixed suspect. We do this to emphasize a point, but the example could be altered to better fit Lindsay *et al.*'s procedure.

[7] The identification probabilities were in fact created by setting the probability of lineup member 3 to a particular value, and randomly generating the probabilities of remaining lineup members. Effective size was set to decrease in direct relation to the identification probability of lineup member 3.

In Figure 1, effective size has five values. For each of these values, the identification probabilities must either average chance expectation (1/6),[8] if there is no 'not present' choice available to witnesses, or ($1-p$(not present)/6) if such a choice is available to witnesses. If the former case holds true, then it follows that the mean probability for each level of effective size in Figure 1 must be 0.167. A plot of effective size against mean identification probability will yield a straight line with equation $y = 0.167$. An attempt to fit a least squares regression line to the unaveraged identification probabilities will yield a line with almost exactly the same equation. Inspection of Figure 1 will show that this is the case, and therefore that the shape of the line shown there was not simply a consequence of choice of hypothetical data.

In the case where witnesses are allowed to choose a 'not-present option' – the recommended way of conducting a lineup in most criminal justice systems – the mean identification probability will not equal chance for every level of effective size, and the mean identification probabilities will not be equivalent across levels of effective size. They will vary as a function only of the number of witnesses who choose the 'not-present' option and the number of lineup members. In the case of the situation depicted by Figure 1, the mean identification probabilities will usually be very close to chance expectation.[9] An attempt to fit a least squares regression line will still yield a line with a near-zero gradient, and will suggest that there is no substantive relationship between effective size and identification probability.

The problem, again, is the statistical dependency introduced by collecting multiple identification probabilities and effective sizes from the same lineups. Under these conditions, it appears that it is not possible to evaluate the predictive capacity of measures of lineup size.

It certainly appears that it is much too early to jettison lineup size as a useful concept, and effective size as a useful measure.[10] It is not clear that effective size fails to predict probability of mistaken identification, since the data offered by Lindsay *et al.* do not provide a test of this ability. What is needed is a set of mock witness experiments which determine independent effective sizes for a number of target-absent lineups, and a set of simulated witness experiments which use these same lineups to estimate independent rates of mistaken identification.

What sample size?

Statistical power has become an important consideration in the behavioral sciences in the last 20 years (Cohen, 1988). Little has been written about statistical power in relation to lineup measures, particularly those that derive from the mock witness procedure, but it is clearly a matter that must be taken seriously.

Although power depends on several factors, the most efficient way in which it can be increased in behavioural research is by paying careful attention to sample size. In the context of the mock witness procedure, this means thinking about the number of mock witnesses, and to a lesser extent, the number of members used in the lineup. Jack Brigham and his associates have suggested that 18 mock witnesses may be a

[8] The probabilities are dependent – they must sum to unity, and since there are six lineup members, the average equals 1/6, which is chance expectation.

[9] If 0.1 or 0.4 or 0.7 of the witnesses choose the 'not-present' option, the mean identification probabilities will be 0.15, 0.1, and 0.05, respectively.

[10] Agresti's estimator, *I*, may be preferable to effective size, though, for statistical reasons.

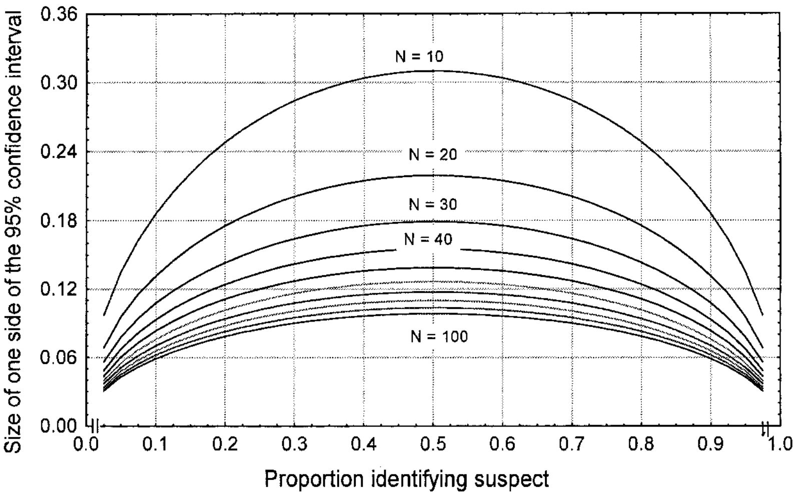

Figure 2. Confidence interval (95%) span as a function of mock witness sample size and proportion of witnesses identifying the suspect

sufficient number for practical purposes (Brigham and Brandt, 1992; Brigham *et al.*, this issue), although Wells and Bradfield (this issue) have been rather more cautious.

We should think about sample size in relation to all the measures of lineup size and lineup bias that researchers have found useful. However, since the most widely used – and least embellished – measure is that of the proportion of mock witnesses who choose the suspect, we can make a beginning by considering the effect of sample size on the confidence interval around this proportion. By implication, this can be extended to one-sample and two-(independent)-sample significance tests.[11]

The confidence interval can be calculated by using the normal approximation to the binomial, although this approximation will be much better for larger samples of mock witnesses. An important fact to note about this confidence interval is that its width will not be constant, but vary in a curvilinear fashion, as a function of the size of the proportion (of mock witnesses identifying the suspect). Figure 2 demonstrates the effect different proportions have on the width of the confidence interval, for different number of mock witnesses. The *Y*-axis, one-sided size of the confidence interval, can be taken as an index of statistical power.

The figure makes clear that larger sample sizes of mock witnesses will substantially reduce the width of the confidence interval for all values that the proportion may take. Larger sample sizes will also reduce the 'amplitude' of the curve, i.e. the varying difference in size of the interval.

For relatively small sample sizes, the confidence interval will be very wide, particularly when the proportion of identifying witnesses is in the neighbourhood of 0.5. Thus, if a researcher uses 20 mock witnesses in an experiment, and 0.38 of these identify the suspect, the one-sided size of the 95% confidence interval around this proportion will be around 0.22 – in other words, the confidence interval will be

[11] In the case of the one sample test, if the $(1-\alpha)*100\%$ confidence interval around the proportion includes the value set by the null hypothesis, then the null hypothesis is accepted. In the case of the two sample test, if the $(1-\alpha)*100\%$ confidence interval around the difference of proportions includes 0, then the null hypothesis is accepted.

approximately (0.16; 0.60). If the confidence interval is used as a proxy for a significance test, we must conclude that the observed proportion is in the range expected by random sampling variation. Increases in the size of the mock witness sample will make the confidence interval a lot narrower, and the significance test a lot more sensitive. Thus, a sample of around 70 mock witnesses will decrease the one-sided size of a confidence interval around a proportion of 0.38 to about 0.1.

It is difficult to make a recommendation regarding an 'ideal' sample size for use in mock witness research. There is a per capita cost in recruiting mock witnesses for an experiment, and this should be weighed up against the desired sensitivity or power of the experiment. It is clear from Figure 2, though, that there are substantial gains in sensitivity to be had by increasing the sample size over the guideline of 20 suggested by Brigham and colleagues.

Interpreting lineup measures

All the lineup measures considered thus far do not explicitly answer whether particular lineups are biased or unfair. They provide quantities which must be interpreted in order to make such decisions. In some of the earliest discussions of measures of lineup bias and lineup size, lineup researchers declared that the actual interpretation of the measures was something best left to the fact finders or the lawyers (Malpass, 1981; Wells *et al.*, 1979).

However, not all eyewitness researchers share this view, and some have suggested criteria for interpreting the measures in terms of fairness and unfairness. In this issue, Brigham *et al.* restate a number of these criteria. For example, the proportion of mock witnesses choosing the suspect is considered to demonstrate bias when it is significantly greater than chance expectation, and similarly, effective size is said to show unfairness when effective size drops to half the nominal size. Brigham *et al.* then proceed to test the criteria empirically against a set of real lineups, gathered in the process of serving as an expert witness in a number of trials, over a number of years. The conclusion Brigham *et al.* draw from their study is that different lineup measures frequently lead to different conclusions regarding the fairness of lineups, and that this makes the task of expert testimony about lineups in court trials somewhat difficult. Lineup measures fail in the sense that they often lead to very different conclusions.

The question of interpretation is a thorny one. Lawyers and Judges will no doubt press expert witnesses to give an opinion on whether a lineup is biased or not, and expert witnesses may therefore feel that they should have some criteria for judging whether particular lineup measures show that a lineup is biased or unfair. However, it may be better not to take over the court's task of deciding whether particular lineups are fair or not. Making decisions about fairness will mean that the expert witness has to set essentially arbitrary thresholds on lineup measures, and setting thresholds of this kind will lead to the kinds of contradiction shown by Brigham *et al.* (this issue).

This is not because the measures say different things about particular lineups, but only because of the way in which threshold criteria force differences between continuous measures. When thresholds are applied to measures of lineup bias and lineup size, bias becomes all-or-nothing (biased versus unbiased), and size becomes all-or-nothing (acceptable versus unacceptable). At the margins of the threshold, near-negligible differences will become amplified to take up endpoints on the new scale. Thus, effective size of 2.9 for a six-person lineup will lead to a conclusion of 'unfair'

Table 3. Correlations between lineup measures reported by Brigham *et al.* for 18 police lineups

	Proportion choosing suspect	Log (functional size)	No. of acceptable lineup members
Log (functional size)	−0.899*		
No. of acceptable lineup members	−0.396*	0.201	
Effective size	−0.544*	0.380*	0.745*

*$p < 0.01$. Correlations are based on a minimum of 44 pairwise cases. This is for more than the 19 lineups evaluated by Brigham *et al.*, since some 'lineups' were evaluated for more than one witness.

whereas effective size of 3.1 for the same lineup will lead to a conclusion of 'fair'. In addition, the use of thresholds overlooks the fact that random sampling variability will be at play when we make estimates of effective size (and other lineup measures). If we use 21 mock witnesses, and the lineup we wish to evaluate has six members, then we can expect a 95% confidence interval around effective size to be at least one unit – i.e. if the effective size estimate is 3.0, the interval around this will be something like (2.3; 3.7).[12] If we follow the advice given by Brigham *et al.*, we would conclude that the lineup size is acceptable, but the large confidence interval suggests greater caution.

Brigham *et al.* (this issue) applied a set of such thresholds to various lineup measures, and concluded that lineup measures lead to very different conclusions in the practical context of deciding on the unfairness of police lineups. However, careful examination of their data confirms that this is a consequence of thresholding criteria, which ensures that lineup measures will frequently disagree with each other when compared on the same lineups. If the measures are allowed to vary continuously, instead of being thresholded, we arrive at a very different conclusion. Table 3 reports correlations between unthresholded lineup measures across 18 lineups, from their study.

It is clear from the Table that all lineup measures show significant correlations with each other, with the exception of the relation between functional size and number of acceptable lineup members. In most instances these correlations are high or very high. It is not the case that these measures lead us to different conclusions, then, so much as the application of arbitrary thresholds that leads us to different conclusions.

REAL AND SIMULATED WITNESSES

All the lineup measures discussed thus far are determined with the mock witness procedure. However, there are two measures which do not assume the mock witness procedure, and are used instead with real or simulated lineups. I want to consider one of these, known as diagnosticity (Wells and Lindsay, 1980; Wells and Turtle, 1986).

DIAGNOSTICITY

Diagnosticity is defined as the amount of potential impact that an identification should have in revising one's opinion of guilt without regard to the prior estimate of

[12] Although there is no known way – as yet – of estimating a confidence interval around ES, there is a way of doing so for a related estimator, Agresti's *I*. The example estimates I give are based on *I*.

guilt – how much more likely the data are to have occurred given the truth of one hypothesis (that the suspect is the criminal) relative to the other (that the suspect is innocent). Diagnosticity is thus a ratio of likelihoods: the ratio of the probability that the suspect is identified given that he is guilty, to the probability that the suspect is identified given that he is innocent.

Diagnosticity has been used in the psychological literature in order to compare the relative success of structural alterations to lineup practice (Lindsay and Wells, 1980; Melara *et al.*, 1989). Valuable as the measure may be, little has been written about its statistical properties. There are several ways of conceptualizing the diagnosticity ratio that allow testing the ratio for statistical significance. I suggest one possibility here.

The diagnosticity ratio is

$$\frac{n_{11}}{n_{+1}} \bigg/ \frac{n_{12}}{n_{+2}},$$

which is a ratio of (estimated) conditional probabilities. It is equivalent to an index widely used in biostatistics, called *relative risk*, since it expresses the probability that a guilty suspect is identified, relative to the risk that an innocent suspect is identified. When diagnosticity is 1.0, the events are equally likely; departure from 1.0, on the other hand, reflects the degree to which events are not equally likely. The measure of relative risk has received a fair amount of statistical research, and there are now several widely accepted inferential methods that can be applied to it. These include (1) constructing confidence intervals around the estimate of relative risk, (2) determining the likelihood that the estimate of relative risk is greater than 1, (3) testing for differences between k independent diagnosticity ratios. For details on these techniques, primary sources are Agresti (1990), and Rothman (1986), and for their application directly to diagnosticity, Tredoux (1998).

New ways of measuring lineup bias and lineup size

In this issue, Lindsay *et al.* suggest a new approach to measuring the 'fairness' of lineups. They make the persuasive argument that lineup measures must be validated against eyewitness identification performance in suspect-absent lineups, and propose a general method of relating mock witness evaluations of lineups to this identification performance. In particular, they propose the use of a simple linear regression equation that predicts the probability of a suspect identification in a suspect-absent lineup from the proportion of mock witnesses who choose the suspect in a mock lineup. They concede that the best approach in the long run will be to cumulate results over many studies, and to determine the weights for the regression equation from the cumulated results, but suggest that researchers in the meanwhile use estimates derived from the set of three studies reported in their paper. Their equation is

$$P = (0.327)M - 0.0042$$

where P = probability of a false positive, M = proportion of mock witnesses. The proposal by Lindsay *et al.* translates a gathered piece of evidence about a lineup into a practical speculative metric, and in this sense it is alluring. However, there are a number of serious problems with such an approach.

I argued earlier that the design used by Lindsay *et al.* introduces a dependency into their data, particularly in respect of the rates of identification of members in target-absent lineups. In particular, suspect identification probabilities calculated for different members of the same lineup must sum to unity.[13] Thus, if lineup member X is falsely identified by a proportion, k, of witnesses, then the maximum identification probability for all other lineup members is $(1 - k)$. This means that the estimates reported for the slope coefficient of the equation will be biased downwards, which can clearly be seen from the equation: the maximum identification probability that the equation is able to predict is 0.3228. In other words, even if every single mock witness identifies a particular lineup member (imagine such a lineup!), the equation tells us that only about one in three eyewitnesses will mistakenly identify that person in a target-absent lineup. Replication of Lindsay *et al.*'s study will in all likelihood show that the estimate of the slope coefficient is a serious underestimate. It should probably not be used at all, even as an interim measure.

But the linear regression approach suggested by Lindsay *et al.* is also not a satisfactory approach, in general terms, i.e. even with independent data. It is well known that regression coefficients are very unstable, and can vary massively from sample to sample. Different replications of Lindsay *et al.*'s design, with independent data, are likely to produce very different estimates of the coefficient. This will be exacerbated by the fact that Lindsay *et al.*'s design involves obtaining identification probabilities from witnesses to staged-crime events. Staging the crime in a different way, with the same lineup members and using the same mock witnesses, may have a profound effect on identification probabilities, substantially altering the regression coefficient. In fact, the equation attempts to generalize a relationship between (1) performance of mock witnesses, and (2) identification performance of real witnesses to lineups involving the same members as appear before the mock witnesses. The problem is that (2) hides a tremendous amount of variation – identification performance of witnesses across different crimes (different levels of illumination, different opportunities for observation, etc.) will certainly vary, and this means that there will be a family of regression equations relating mock witness performance to eyewitness identification performance in different staged crimes, but with the same lineup members. One regression equation cannot suffice, unless we assume that there is no variation between identification performance across different crimes.

Finally, the regression equation approach uses only the proportion of identifying mock witnesses as a predictor, and this overlooks the size of the lineups that mock witnesses have been asked to evaluate. Thus, if 1/2 of mock witnesses choose a particular lineup member from a two-person lineup, this is given the same weight as the situation where 1/2 of mock witnesses choose a particular lineup member from a 20-person lineup, and will generate the same probability estimate for a false positive. In the former case, though, performance of mock witnesses is at chance levels, and in the latter case, performance of mock witnesses is massively above chance. This seems clearly mistaken, and underscores the importance of the notion of lineup size.

Lindsay *et al.* make a valuable point when they state that measures of lineup bias and lineup size need to be validated. However, it is not clear that the translation of this observation into a predictive equation is of any value – and it certainly introduces

[13] This depends on whether the no-choice option is used in the simulated lineup. If it is not used, then the sum is unity, but if it is not, then the sum is $(1 - p(\text{no choice}))$.

significant problems. It may be enough to point out to the jury that, as far as we know, mock witness identifications are fairly strongly related to probability of mistaken identifications from perpetrator-absent lineups.

DISCUSSION

There is little in the psychological literature regarding the application of inferential statistical reasoning to lineup tasks and measures. I have argued that this is a failing, and I have suggested some ways in this paper – and an earlier paper (Tredoux, 1998) – in which researchers and practitioners can start to remedy this. Without taking inferential considerations into account, there are problems for both research and application.

The problems are particularly acute in applied legal settings, where measures of lineup bias and lineup size have been used by expert witnesses to support fairness evaluations of police lineups (see Buckhout *et al.*, 1988). If an expert reports in court that 33% of mock witnesses identified a suspect, and that the functional size of the lineup is therefore three, without giving an idea of the inherent sampling variability around the estimate, the testimony may be misleading. If the estimate is derived from a mock witness task in which there are comparatively few mock witnesses, then the sampling variation around this estimate will be fairly great, and a point estimate will not reveal this to the court. This is an important point, because some researchers, including Brigham *et al.* in this issue, have made recommendations that as few as eighteen mock witnesses be used to derive lineup indices. I have argued in this paper that such small numbers of mock witnesses will, on the contrary, lead to lineup estimates that are subject to a great deal of sampling variability, and will therefore be statistically unreliable.

Inattention to statistical properties of the mock witness paradigm and other lineup tasks also leads, in my opinion, to untimely calls from Lindsay *et al.* (this issue) that we discard the concept of lineup size, and the measures of lineup size proposed by Malpass (1981). They argue that measures of lineup size do not postdict eyewitness identification accuracy, but the design they use to do this introduces statistical dependencies that make this conclusion inevitable (and artefactual). The general point they make about the need to validate lineup measures against eyewitness performance is important, but the research they report in this respect does not provide useful data on the issue. It remains for research designs that are not tempted by the luxury of *k* data points per *k* member lineup to gather relevant data.

Lindsay *et al.* take the important point about the need for validation of lineup measures one step further than arguing for the 'end of lineup size', and make a suggestion for a new measure of lineup fairness. In particular, they suggest that a linear regression equation be used to predict false positives from the number of mock witness choices of the suspect. They go further and suggest that estimates of the equation parameters be derived from the largest possible data set, which happens at this stage to be theirs. I have argued that the parameter estimates should not be based on their data set, due to the multiple statistical dependencies that are a consequence of their research design. However, there is also a strong case against the general approach of using a linear regression equation in the suggested way, and I have argued at some length in this paper against it.

Lineup measures are ultimately interesting because they make it possible to apply social science research to a real-world problem. They often need to be 'translated' into the world of juries and courtrooms, though, to have direct application. The problem of 'translation' or 'interpretation' is one that has puzzled eyewitness researchers for two decades. Suggestions have mostly been to avoid interpreting lineup measures (e.g. deciding what value of functional size marks a lineup as unfair), but Brigham and his associates (Brigham *et al.*, 1990; Brigham and Brandt, 1992; Brigham and Pfeiffer, 1994) have made several suggestions regarding cut-off or threshold points demarcating acceptable and unacceptable levels of lineup bias and lineup size. In this paper they report data from 18 real lineups where the use of these thresholds leads to inconsistent conclusions regarding the fairness and/or bias of the lineups, and they warn of the consequences of presenting evidence of this kind to courts.

It is clear that expert witnesses are constantly asked in court to bridge the gap between theory and practice, and that experts who testify in eyewitness cases are no exception. However, it is not a solution to decide on threshold values for lineup measures for the mere sake of building such a bridge. Indeed, I have argued that the use of threshold values creates the interpretative problems identified by Brigham *et al.*, rather than reflecting the difficulty of interpreting lineup measures. Analysis of the unthresholded lineup measures reported by Brigham *et al.* suggests that the measures provide consistent information about the quality of the eighteen lineups.

It is not time to jettison lineup measures. On the contrary, work around the estimation of lineup fairness remains one of the most significant achievements of eyewitness research, and should continue to attract our attention.

REFERENCES

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Agresti, A. and Agresti, B. (1978). Statistical analysis of qualitative variation. In K. F. Schuessler (Ed.), *Sociological methodology* (pp. 204–237). San Francisco: Jossey-Bass.

Brigham, J. C. and Brandt, C. C. (1992). Measuring lineup fairness: Mock witness responses versus direct evaluations of lineups. *Law and Human Behavior*, **16**, 475–489.

Brigham, J. C. and Pfeiffer, J. E. (1994). Evaluating lineup fairness. In D. F. Ross, J. D. Read and M. P. Toglia (Eds.), *Adult eyewitness testimony* (pp. 201–222). New York: Cambridge University Press.

Brigham, J. C., Ready, D. J. and Speir, S. A. (1990). Standards for evaluating the fairness of photograph lineups. *Basic and Applied Social Psychology*, **11**, 149–163.

Buekhout, R., Rabinowitz, M., Alfonso, V., Kanellis, D. and Anderson, J. (1988). Empirical assessment of lineups. *Law and Human Behavior*, **12**(3), 323–331.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Doob, A. N. and Kirshenbaum, H. M. (1973). Bias in police lineups – partial remembering. *Journal of Police Science and Administration*, **1**, 287–293.

Hays, W. (1994). *Statistics*. New York: Holt, Rhinehart and Winston.

Lindsay, R. C. L. and Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior*, **4**, 303–313.

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior*, **5**(4), 299–309.

Malpass, R. S. and Devine, P. G. (1983). Measuring the fairness of eyewitness identification lineups. In S. M. A. Lloyd-Bostock and B. R. Clifford (Eds.), *Evaluating witness evidence* (pp. 81–102). London: Wiley.

Melara, R. D., De Witt-Rickards, T. and O'Brien, T. P. (1989). Enhancing lineup identification accuracy: two codes are better than one. *Journal of Applied Psychology*, **74**, 706–713.

Rothman, K. J. (1986). *Modern epidemiology*. Boston, MA: Little, Brown.

Tredoux, C. G. (1998). Statistical inference on lineup measures. *Law and Human Behavior*, **22**(2), 217–237.

Wells, G. L., Leippe, M. R. and Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior*, **3**, 285–293.

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

Wells, G. L. and Turtle, J. W. (1986). Eyewitness identification: the importance of lineup models. *Psychological Bulletin*, **99**, 320–329.