Law and Human Behavior, Vol. 3, No. 4, 1979

Notes/Discussions

Guidelines for Empirically Assessing the Fairness of a Lineup*

Gary L. Wells,† Michael R. Leippe,‡ and Thomas M. Ostrom§

Issues regarding the fairness of lineups used for criminal identification are discussed in the context of a distinction between nominal size and functional size. Nominal size (the number of persons in the lineup) is less important for determining the fairness of a lineup than is functional size (the number of lineup members resembling the criminal). Functional size decreases to the extent that the nonsuspect members of the lineup are easily ruled out as not being suspected by the police. The extent to which the identification of the suspect can be considered an independently derived piece of incriminating evidence is positively related to functional size. Empirical estimates of functional size can be obtained through pictures of the corporal lineup from which mock witnesses make guesses of whom they believe the police suspect. A distinction is made between a functional size approach and hypothesis testing approaches. Uses of functional size notions in the court, by police, and in research are discussed.

INTRODUCTION

The identification of criminal suspects by means of pictures or corporal lineups has been generally regarded by potential jurors, law enforcement officials, and judges as among the most persuasive incriminating evidence that can be presented (Wall, 1965). "Identifying the defendant as the wrongdoer presents an issue, and often the sole one

*The authors would like to thank Michael Baumgander, Rodney Bassett, Nehemia Geva, John Lingle, Richard E. Petty, and David Simpson for aid in collecting data on the U.S. v. Mills case and Tamara Ferguson, R. C. L. Lindsay, and Anthony G. Greenwald for comments on an earlier version of this manuscript. Portions of the U.S. vs. Mills data were presented at the American Psychological Association meeting in Washington, D.C. in 1976. Requests for reprints should be sent to Gary L. Wells, Department of Psychology, University of Alberta, Edmonton, Alberta, T6G 2E9.

†Department of Psychology, University of Alberta, Canada.
‡Department of Psychology, St. Norbert College, Wisconsin.
§Department of Psychology, Ohio State University.

0147-7307/79/1200-0285$03.00/0

for determination, in every criminal trial" (Woocher, 1977, p. 961). Experimental investigations of eyewitness performance, however, tend to show discouragingly high rates of error (e.g., Buckhout, 1974). In addition, numerous examples of misidentification in criminal cases have been documented (see Borchard, 1932; Frank, 1949; Marshall, 1969) and "inaccurate identification has been and continues to be a major source of faulty convictions" (Levine and Tapp, 1973).

One source of this error can be within the structure of the lineup or picture array from which the defendant was identified by the witness(es). Using staged-crime research, Buckhout (1974) has shown that witnesses are more likely to choose an innocent member of a picture array, and presumably a lineup, if he or she is distinctive from the other members (e.g., if his or her picture is taken at a slightly different angle). The defendant's picture is often distinctive in picture arrays since it may come from any source while the nonsuspects' pictures generally come from police files that are fairly standard in format, quality, lighting, and so on. In corporal lineups the defendant might dress differently or stand differently from the other lineup members who are frequently police detectives (Levine & Tapp, 1973) and, therefore, are more homogeneous in appearance, posture, and general demeanor.

There is little doubt that lineups can be biased so as to increase the chances that the suspect will be identified. As outlined elsewhere (e.g., Buckhout, 1974; Doob & Kirshenbaum, 1973; Levine & Tapp, 1973), witnesses can be characterized as (a) wanting to please police investigators (b) presumptive that the police have good evidence against one of the lineup members, and (c) seeking of cues regarding which response is appropriate (i.e., which lineup member to choose). Recent empirical evidence further shows that the witness may have little knowledge of the extent to which she or he is making a false identification. Specifically, staged crime experiments by Leippe, Wells, and Ostrom (1978) and Wells, Lindsay, and Ferguson (1979) have shown that a witness who makes a false identification is not appreciably less confident than is a witness who makes an accurate identification. Thus, witnesses show little or no awareness of the factors that may lead to faulty identifications.

The criminal justice system treats an identification of a suspect from a lineup as an independent piece of evidence. Strictly speaking, of course, it is not totally independent evidence since the fact that a specific defendant appeared in the lineup was dependent on prior evidence, perhaps circumstantial, which made the investigators suspect that particular defendant. The independence issue, however, is a matter of degree and the structure of the corporal lineup can affect the degree to which the identification process is independent of the police investigators' prior suspicions. In general, to the extent that the witness can discern whom the police suspect in the lineup, the witness's identification is losing independence from the police investigators' prior suspicions.

Lineups have a special status in eyewitness identification research. As Wells (1978) outlined, lineups are one of the few variables that can be characterized as *system* variables. That is, unlike some variables that affect the rate of eyewitness identification errors (e.g., cross versus same race identifications, see Brigham and Barkowitz, 1978; Elliott, Wills, and Goldstein, 1973; Malpass and Kravitz, 1969), lineup structure can be controlled by the criminal justice system.

There are no definitive guidelines for the construction of a corporal lineup. Some eyewitness researchers have attempted to list potential sources of lineup bias (e.g., Levine and Tapp, 1973; Fanslow, 1975). These include such factors as how many peo-

ple are in the lineup, whether the lineup participants are mixed in race or skin tone, have different amounts and/or styles of facial hair, vary greatly in age, height, modes of dress, body frames, and so on (see Katz et al., 1975). However, one of the problems with a checklist-type approach to factors regarding the fairness of a lineup is that the various factors are multidimensional and have unknown weights. Thus, it is unclear how a psychologist could summarize the fairness of the lineup by using a checklist approach. There is also some question regarding the extent to which interjudge reliability is sufficient to allow only one psychologist to make a determination regarding the fairness of a lineup or whether several psychologists should combine their judgments. Finally, psychologists have a tendency to develop a "keen eye" for their area of expertise and are quite likely to identify biasing factors in the lineup that would go unnoticed by the lay witness.

As far as we can determine, the first attempt to empirically assess the fairness of a lineup was done by Doob and Kirshenbaum (1973). Doob and Kirshenbaum reported the study of two criminal cases for which they have photographs of the corporal lineups from which the defendants were identified. Doob and Kirshenbaum defined a biased lineup as "one where a person who was not a witness to the crime is more likely to pick the suspect out of the lineup than we would expect by chance (where chance is defined a $1/n$, n being the number of people in the lineup)" (Doob & Kirshenbaum, 1973, p. 290). Doob & Kirshenbaum's seminal use of a mock-witness paradigm, while seemingly simple, deserves considerable credit. It avoids the "checklist-approach" problems listed above and provides a framework for further development. The current article is an attempt to improve and elaborate on Doob and Kirshenbaum's approach. In doing so, certain shortcomings of their approach must be noted.

First, consider the case of a 6-person lineup. Suppose that the lineup as considered fair by Doob and Kirshenbaum's criteria had the proportion of mock witnesses' choosing the defendant as 1/6. Further suppose that foils are added to the lineup who in no way resemble the criminal so that there are the original six lineup members plus 5 bad foils for a total of 11 people. Suddenly we would discover that the lineup is biased because the proportion of mock witnesses choosing the defendant would remain at 1/6 (bad foils draw no choices) while chance is considered 1/11. Yet, as Doob and Kirshenbaum acknowledge, the critical issue concerns the extent to which there are sufficient distractors to not bias the witness toward the suspect. What is needed is summary statistic that will show that these two lineups are equally fair.

Second, consider a case wherein there is only one person in the lineup, namely the suspect. This state of affairs is termed a show-up and has been generally considered illegal (Stovall v. Denno, 1967). The Doob and Kirshenbaum approach, however, is totally insensitive to detecting the unfairness of show-ups because there is no way for the number of mock-witness choices of the suspect to exceed $1/n$. Further consider what happens if we add more foils to the one-person lineup. This time, however, we add good *foils* so that choices of the suspect remain at $1/n$. The Doob and Kirshenbaum approach will fail to label the additional foils as being a benefit to the fairness of the lineup. Thus, what is needed is summary statistic that will sensitively reflect such a change in the characteristics of the lineup.

We propose using the probability of mock witnesses choosing the suspect as the summary statistic rather than testing the probabilities against a null hypothesis of

$1/n$. Thus, the relevant statistic describing a lineup is D/n, where D is the number of mock witnesses choosing the defendant and n is the number of mock witnesses. We also personally prefer a reciprocal transformation of this probability into n/D which reflects what we will term the "functional size" of a lineup. That is, if the number of mock witnesses choosing the suspect is 50% of all mock witnesses then the probability (i.e., D/n) is .50 and the functional size (i.e., n/D) is 2.0. Our preference for functional size rather than probability is largely because of the imagery connoted by functional size since it reflects the number of feasible lineup members. Our interactions with legal practitioners (albeit somewhat anecdotal) leads us to conclude that a functional size concept is more readily grasped by such practitioners than is probabilities. (Readers can transform functional size into probabilities if they wish since the conclusions do not change.)

How does the concept of functional size differ from Doob & Kirshenbaum's (1973) approach? Consider the earlier stated problem with Doob & Kirshenbaum's analysis in which a fair six-person lineup has bad foils added to it. Doob & Kirshenbaum's analysis would, at some point, label the lineup as biased since choices of the defendant would no longer match $1/n$. The functional size approach, however, will continue to give the same estimate of functional size (i.e., functional size = 6.0) regardless of how many bad foils are added. Thus, the functional size estimate is invariant and appropriately insensitive to the irrelevant addition of lineup members. That is, the functional-size estimate appropriately acknowledges that these bad foils are functionally not there. Obviously, adding a bad foil is like adding a dog or refrigerator to the lineup—if it draws no choices from the mock witnesses it is functionally irrelevant to the resultant statistic. Note, however, that if good foils are added to this lineup (i.e., foils that draw choices) then the functional size value appropriately reflects such additions and shows the resultant lineup (e.g., functional size increases to 11.0) to be even fairer to the defendant. The best that the Doob & Kirshenbaum approach can do, however, is consider the addition of five good foils to be, at best, simply maintaining a fair lineup.

Considering the case of a show-up, the functional size approach would appropriately label a show-up as such (i.e., functional size = 1.0). The functional size approach clearly distinguishes a show-up from a six-person lineup with six good foils (i.e., functional size = 1.0 versus 6.0) whereas the Doob & Kirshenbaum approach labels each as equally fair.

STATISTICAL INFERENCES REGARDING FUNCTIONAL SIZE

An obtained functional size can, of course, be misleading if the number of mock witnesses is small. At the extreme, one can imagine a 6-person lineup without true biases (i.e., functional size also equals 6) in which only one mock witness is used. In such a case, there would be a one-sixth chance of obtaining a functional size of one and a five-sixth chance of obtaining a functional size of infinity. The fact that as the number of mock witnesses increases, the obtained functional size estimate approximates the actual functional size is simply another example of Bernouli's theorem. The critical question is: How many mock witnesses are necessary to test the functional size of a lineup or picture array?

There is no definitive answer to the question of sample size. However, there are certain guidelines that seem reasonable. We suggest that an investigator utilize enough mock witnesses to produce a 95% confidence limit of ± .5 on the functional size. Thus, an investigator who finds a functional size of 3 could assert with 95% confidence that the true functional size is closer to 3 than it is to 2 or 4.

However, it should be noted that the number of mock witnesses required to establish a 95% confidence interval of ± .5 on functional size is not a constant number. Instead, the required sample size depends on the obtained sample mean. Essentially, there are two factors operating. First, because the distribution of data is binomial (the mock witness either chooses the defendant or not), the greatest variance is associated with the case in which one-half of the mock witnesses choose the defendant. Therefore, sample variance is greatest when the obtained functional size is 2.0 and variance is lower when the functional size is less than or greater than 2.0. Secondly, however, functional size is a reciprocal data transformation of the proportion choosing the defendant. Therefore, the proportional difference between the functional sizes of 5.0 and 5.5 (2%) is smaller than the proportional difference between functional sizes of 2.0 and 2.5 (10%). Overall, it turns out that the number of mock witnesses required to establish a 95% confidence interval of ± .5 around the obtained functional size value is a monotonically increasing function of the obtained functional size.

It may first appear that the sample size issue is more problematic for the functional size approach than it is for Doob & Kirshenbaum's (1973) approach. The fact is that sample size is a tough issue for each approach but, we argue, it is actually less problematic for the functional size approach. To understand why this is the case, it must be noted that Doob & Kirshenbaum's approach is a hypothesis-testing procedure whereas functional size is a parameter-estimation procedure. That is, Doob & Kirshenbaum's analysis calls for testing an obtained value against a null hypothesis value of $1/n$ and using some criterion for the resultant t-test (e.g., .05 level of significance). Because the distribution is binomial, the probability of obtaining a value significantly above $1/n$ (i.e., a biased lineup) is a function of sample size and the magnitude of difference between $1/n$ and the obtained value. Thus, sample size is an important determinant of whether or not a given lineup will be judged biased with Doob & Kirshenbaum's procedure. Doob & Kirshenbaum did not suggest a particular sample size for their approach because in doing so *they would in effect be specifying a particular difference between an obtained value and* $1/n$ *that would be called a biased lineup.* In other words, specifying sample size is a value judgment for Doob & Kirshenbaum's approach because it totally determines how far above $1/n$ an obtained value must be to be called biased. While increasing sample size increases the chances that a lineup will be declared biased using Doob & Kirshenbaum's approach, increasing sample size simply decreases the size of the confidence limits around a given functional size using the functional size approach. The current authors were able to specify a sample size requirement because sample size does not determine whether a lineup is biased or not using the functional size approach. The functional size approach leaves fairness or bias decisions with the courts and increasing sample size simply increases the stability or certainty associated with the estimated parameter. Our recommendation that an obtained functional size (e.g., 3.0) have a sufficient sample size to place the nearest whole numbers (2.0 and 4.0) outside of its 95% confidence limits is not a statement related to fairness or bias criteria. No addition to the sample

Table 1. Percentage of Mock Witnesses Choosing Each Lineup Member as the Accused

	#1	#2	#3	#4	#5	#6 (Mills)
Percentage of mock witnesses indicating member as accused	4.9%	2.4%	4.9%	19.5%	7.3%	60.9%

size will increase the chances of the lineup being declared biased by the scientist because the functional size approach is a parameter-estimation rather than hypothesis-testing approach.

U.S. V. MILLS: AN EMPIRICAL EXAMPLE

We employed the mock-witness paradigm in the case of U.S. v. Mills. In that particular case, Mills was accused of bank robbery and our involvement was solicited by the defense. The bank robber was described by three witnesses as "black, male, short, full beard, and thin but not skinny." We gave 60 subjects (introductory psychology students) this general description, a description of the crime and a picture of the actual corporal lineup (6 members) from which Mills was identified. The mock witnesses (subjects) were asked to pick the person they thought to be the accused and were allowed to pick "none of the above." The mock witnesses were tested individually. Of the 60 mock witnesses, 41 made a choice and 19 indicated "none of the above."[1] Table 1 presents the frequencies with which the various lineup members were picked by the mock witnesses. Mills was the most frequent choice with 61% of the mock witnesses who made a choice believing that he was the lineup's true defendant. The data indicate that the best guess regarding the true functional size of the lineup is that it was composed of only 1.64 members.

Applying our confidence limit criterion from above, we can conclude with 95% confidence that the true sample-mean functional size is between 1.31 and 2.19.[2] Note that these upper and lower confidence limits are not equal distances from the sample mean of 1.64. This is because the confidence interval is calculated in the form of proportions while functional size is a *reciprocal* transformation of proportions and is, therefore, skewed. The nature of this skew follows a pattern wherein the upper limit of

[1]We did not force choices and therefore must exclude these 19 mock witnesses from the choice analysis. However, we did obtain likelihood estimates from these 19 persons. That is, each mock witness assigned a number to each lineup member indicating the likelihood that each lineup member was the suspect. Assuming that the highest likelihood is the lineup member who would have been chosen had we forced the mock witnesses to make a choice, the 19 mock witnesses followed the same general pattern as the other 41 mock witnesses. Specifically, Mills was associated with the greatest likelihood estimate in 13 (68.4%) of the 19 cases.

[2]This test was based on 40 degrees of freedom. The confidence limits were calculated from a normal t-test approximation to the binomial. In general, the t-test normal approxmation can be used for binomial data whenever the smaller Np (number of mock witnesses times the proportion who choose the defendant) or Nq (number of mock witnesses times the proportion who do *not* choose the defendant) is ≥ 10 (see Hays, 1973, p. 230).

the 95% confidence interval around an obtained functional size will always be a greater distance from the obtained functional size than will the lower limit. Thus, a conservative approach would be to ensure that the *upper* limit of the 95% confidence interval does not exceed the obtained functional size by more than approximately + .5.

CONCLUSION

A technique was outlined to assess the fairness of a lineup or picture array. While the technique empirically assesses the extent of certain biases in identification procedures, it leaves the question of "fairness" to the courts by simply specifying functional size. It should be noted, however, that while we feel it is up to the courts to decide whether or not a particular sized lineup is fair, there is a well-understood precedence for considering a "show-up" to be unacceptable (Stovall v. Denno, 1967). A show-up is a one-to-one confrontation between the witness and a suspect. It can be argued that a functional size of 1.0 is essentially a show-up. The future question for the courts, then, is to decide whether a lineup with a functional size of 2 or 3 or any other low number might be judged as too close to a show-up to be acceptable. The arguments presented in Stovall v. Denno (1967) regarding high risks of suggestion and pressure that result from show-ups may apply in varying degrees to lineups with low functional sizes.

Because the functional size of a lineup is simply a reciprocal of a probability, why not report probabilities? When one is analyzing a lineup for purposes of presentation to researchers, probabilities are perhaps a superior mode of presentation. However, for either the presentation of information to lawyers or for courtroom exposition to jurors and judges, it is far easier to understand that a 6-person lineup is functionally composed of only 3 persons than to explain the notion of a .33 probability of the defendant being chosen. Researchers in human decision making have extensively documented how the nonstatistician's use of probabilities represents a fundamental misunderstanding of probability principles (see Einhorn & Hogarth, 1978).

Current courtroom discussions of lineup fairness are subjective and confusing, with no empirical or scientific base. Law-enforcement officials have not been provided with specific criteria by which they can assess their own procedures, and the logical issue of a fair lineup is often phrased in confusing, ill-defined verbal statements. The functional size approach presents a standard procedure for obtaining objective information about a lineup. Adopting this standard approach means that the court need only be given the obtained functional size estimate because assurances about reasonable statistical reliability are already built into the computational system via the sample-size guidelines above.

Extreme examples of cases where functional size is small are not uncommon, such as in State v. Parker (1969), wherein the victim alleged he had been assaulted by 3 Indians and the 3 suspects were the only 3 Indians in a six-man lineup. More subtle biases are exemplified in State v. Burch (1969), where the defendant differed markedly from the other lineup members in complexion; Massen v. State (1969), where the other lineup members did not match the suspect's hair color; People v. Chambers (1969), where the defendant was clothed differently from the other lineup members; and People v. Stanton (1969), where the suspect was actually *directed* by police to

dress in a particular manner for the lineup identification task. But, most cases in which functional sizes are small are probably less obvious than these because it is often many factors (rather than a simple factor such as complexion) that cut away at nominal size to yield a low functional size.

There are two places where the functional size concept can fit into the criminal justice process. On the one hand, the defense and/or prosecution agents can hire behavioral scientists to calculate the functional size of a lineup on a post hoc basis using the corporate-lineup picture. Alternatively, a somewhat better procedure would be for the police investigators to specifically construct lineups that meet reasonable *functional* size requirements as opposed to a concern only for the number of lineup members. We believe that police do *not* need to use subjects, empirically assess functional size, and determine confidence limits. *Empirical* estimates of functional size need only be an objective "check" on the sincerity of police investigators. Conscientious attention to lineup construction and knowledge of potential biases on the part of police investigators should be sufficient. In other words, a lineup with a large functional size can be easily constructed without resort to empirical testing. Hopefully, the police investigators' knowledge that an empirical test may be conducted would be enough to produce lineups with reasonable functional sizes.

The functional size technique outlined herein should apply not only to actual police lineups, but, also to eyewitness-identification research. A staged "crime" is an important research tool in studying eyewitness identification, because only when crimes are independently manipulated can the investigator make categorical assessments of whether a witness was or was not correct. However, only two staged-crime experiments have ever specified the functional size of their lineups (Leippe, Wells, & Ostrom, 1978; Wells, Lindsay & Ferguson, 1979). As Wells (1978) noted, it is extremely difficult to make comparisons of accuracy rates across experiments because it is not clear whether any observed differences are due to the nature of the witnessed event or due to the nature of the subsequent lineup. To attain a higher level of replicability of results in staged-crime research, the published report should specify the functional size of the lineup or picture array that was used.

Finally, recent evidence indicates that jurors cannot detect differences between accurate and false-identification witnesses (Wells, Lindsay, and Ferguson, 1979). Therefore, we must rely on the ability of police authorities to at least partially screen out inaccurate witnesses through testing procedures, since jurors are as likely to believe inaccurate witnesses as they are to believe accurate ones.

REFERENCES

Borchard, E. M. *Convicting the innocent: Errors of criminal justice.* New Haven: Yale, 1932.

Brigham, J.C., and Barkowitz, P. Do they all look alike? Experience, attitudes, and the ability to recognize faces. *Journal of Applied Social Psychology,* 1978, **8,** 306–318.

Buckhout, R. Eyewitness testimony. *Scientific American,* 1974, **321,** 23–31.

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

Einhorn, H. J., & Hogarth, R. M. Confidence in judgment: Persistence of the illusion of validity. *Psychological Review,* 1978, **85,** 395–416.

Elliot, E.S., Wills, E. J., & Goldstein, A. G. The effects of discrimination training on the recognition of white and oriental faces. *Bulletin of the Psychonomic Society,* 1973, **2,** 71–73.

Fanslow, M. How to bias an eyewitness: A review of research on expectancy applied to eyewitness identification testing. *Social Action and the Law Newsletter,* 1975, **2,** No. 3, 3–6.

Frank, J. *Courts on trial.* Princeton: Princeton University, 1949.

Hays, W. L. *Statistics for the social sciences.* New York: Holt, Rineheart, & Winston, 1973.

Katz, R., Vesel, B., Buckhout, R., & Wolft (Eds.) A reliability checklist for lineups. *Social Action and the Law Newsletter,* 1975, **2,** No. 3., 9–10.

Leippe, M. R., Wells, G. L., & Ostrom, T. M. Crime seriousness as a determinant of accuracy in eyewitness identification. *Journal of Applied Psychology,* 1978, **63,** 345–351.

Levine, F., & Tapp, J. The psychology of criminal identification: The gap from Wade to Kirby. *University of Pennsylvania Law Review,* 1973, **5,** 1079–1131.

Malpass, R. S., & Kravitz, J. Recognition for faces of own and other race. *Journal of Personality and Social Psychology,* 1969, **13,** 330–334.

Marshall, J., *Law and psychology in conflict.* New York: Doubleday-Anchor, 1969.

Massen v. State, 41 Wis. 2d 245, 1969.

People v. Chambers, 112 Ill. App. 2d 347, 1969.

People v. Stanton, 274 Cal. App. 2d 13, 1969.

State v. Burch, 284 Minn. 300, 1969.

State v. Parker, 282 Minn. 343, 1969.

Stovall v. Denno, 388 U.S. 263, 1967.

Wall, P. *Eyewitness identification in criminal cases.* Springfield Illinois: Charles C. Thomas, 1965.

Wells, G. L. Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology,* 1978, **36,** 1546–1557.

Wells, G. L., Lindsay, R.C.L. & Ferguson, T. Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology,* 1979, **64,** 440–448.

Woocher, F. D. Did your eyes deceive you? Expert psychological testimony on the unreliability of eyewitness identification. *Stanford Law Review,* 1977, **29,** 969–1030.