

Regularities in Eyewitness Identification

Steven E. Clark · Ryan T. Howell · Sherrie L. Davey

Published online: 5 April 2007

© American Psychology-Law Society/Division 41 of the American Psychological Association 2007

Abstract What do eyewitness identification experiments typically show? We address this question through a meta-analysis of 94 comparisons between target-present and target-absent lineups. The analyses showed that: (a) correct identifications and correct-nonidentifications were uncorrelated, (b) suspect identifications were more diagnostic with respect to the suspect's guilt or innocence than any other response, (c) nonidentifications were diagnostic of the suspect's innocence, (d) the diagnosticity of foil identifications depended on lineup composition, and (e) don't know responses were nondiagnostic with respect to guilt or innocence. Results of diagnosticity analyses for simultaneous and sequential lineups varied for full-sample versus direct-comparison analyses. Diagnosticity patterns also varied as a function of lineup composition. Theoretical, forensic, and legal implications are discussed.

Keywords Eyewitness identification · Meta-analysis

This paper addresses a simple question: What do the results of eyewitness identification experiments typically show? It is a question that appears straightforward and simple enough, and yet is difficult to answer.

For example, given that the research has been concerned for decades with the reliability of eyewitness testimony, one might ask, *is eyewitness identification evidence reliable?* One way to address this question might be to look at the correct identification rates from eyewitness identification experiments. In doing so, one will discover quickly that these correct identification rates vary widely across experiments, as high as 80% (Pozzulo & Lindsay, 1999) and as low as 8% (Parker & Ryan, 1993). The variability in experimental outcomes has led some researchers to question whether the experimental literature is of much help in informing the court (see Ebbesen & Konecni, 1996; Egeth, 1993, 1995; Elliott, 1993; McCloskey & Egeth, 1983). Of course, variability in results *should* arise from variation in experimental procedures. However, even in

S. E. Clark (✉) · S. L. Davey
Psychology Department, University of California, Riverside, CA 92521, USA
e-mail: clark@ucr.edu

R. T. Howell
California State University, Bakersfield, CA, USA

the midst of such variability, there should be consistent *patterns* of responding, and regularities that appear consistently across experiments.

Meta-analyses have begun to appear in the last decade that serve the purpose of standing back from the trees to get the wider picture of the forest, allowing the field to observe relatively consistent patterns of responding with respect to stress (Deffenbacher, Bornstein, Penrod, & McGorty, 2004), prelineup mugshot exposure (Deffenbacher, Bornstein, & Penrod, 2006) weapon focus (Stebly, 1992), simultaneous and sequential lineups (Stebly, Dysart, Fulero, & Lindsay, 2001), instructional bias (Clark, 2005; Stebly, 1997), and comparisons of showups and lineups (Stebly, Dysart, Fulero, & Lindsay, 2003).

The present paper does not address the issue of regularity with a discussion of a list of various reliable effects, but rather, by looking at very general patterns of responding. We suggest that the most important, and most often-manipulated, variable in the eyewitness identification research literature is the presence or absence of the target in the lineup.

The importance of comparing target-present and target-absent lineups

In a real criminal investigation the person suspected of having committed the crime may be guilty of the crime, but may also be innocent. These two possibilities are simulated in laboratory experiments by presenting participant-witnesses with target-present (TP) lineups, which simulate the case in which the suspect is guilty, and target-absent (TA) lineups, which simulate the case in which the suspect is innocent. The importance of considering both conditions has been discussed repeatedly in the literature (Clark & Tunnicliff, 2001; Wells, 1993; Wells & Turtle, 1986), and it is not necessary to review those arguments in any great detail here.

The question of guilt versus innocence is typically viewed as a component to an interaction effect. Specifically, the interpretation of the results from one lineup condition depends on the outcome of the other lineup condition. For example, a new technique that results in more correct target identifications in TP lineups might be viewed quite favorably, unless it were shown that the technique also increases the likelihood of false identifications in TA lineups.

The comparison of TP and TA lineups is more than a factor to consider in evaluating other variables. Rather, the comparison of TP and TA lineups is a measure of the “guilt effect,” or the “innocence effect,” depending on one’s viewpoint. The main effect question is: How do witnesses respond to the lineup when the suspect is guilty versus when the suspect is innocent?

The importance of considering target presence versus absence as a within-experiment variable was noted by the New York Supreme Court (*People v. Smith*, 2004) as it denied a defense motion to admit expert testimony on eyewitness evidence: “. . . much of the research directed to lineup procedures uses either (usually) target present or (rarely) target absent data, but not both.” The present paper analyzes the data from 94 direct comparisons between target-present and target-absent conditions, suggesting that the New York Court was incorrect in its assessment of the eyewitness identification research literature.

Outcomes of eyewitness identification experiments

Table 1 shows the various response outcomes in typical eyewitness identification experiments. These experiments categorize each response as an identification of the suspect, an identification of a foil, or a nonidentification. In a small subset of the studies reviewed here, the nonidentification responses could be further divided into don’t know and reject responses. Since only a few studies distinguish between don’t know and reject responses, most of our analyses will use nonidentification responses which collapse over these two responses for studies that make the distinction. We will use the abbreviations in Table 1 throughout this paper. Thus, susTP refers

Table 1 Response outcomes for target-present and target absent lineups

	Target	
	Present	Absent
Suspect	suspTP	suspTA
Foil	foilTP	foilTA
Nonidentification	noidTP	noidTA
Don't know	dkTP	dkTA
Reject	rejTP	rejTA

Note. susp, foil, noid, dk, and rej denote suspect, foil, nonidentification, don't know, and reject responses; TP and TA denote target-present and target-absent lineup conditions.

to a suspect identification in a TP lineup, suspTA refers to a suspect identification in a TA lineup, foilTP refers to a foil identification in a TP lineup, and so on.

Many published studies do not distinguish in TA lineups between the identification of an innocent suspect (suspTA) versus the identification of a foil (foilTA). Of course, these two responses are conceptually and forensically quite different. Nonetheless, many studies report any identification made for a target-absent lineup as a foil identification. For these studies, the TA identifications were separated to give estimates for suspect and foil identifications. Specifically, the suspect identification rate was calculated by dividing the identification rate by K the lineup size, and the foil identification rate was estimated as $K-1$ times the suspect identification rate. This method of estimating separate suspect and foil identification rates in TA lineups implies that the innocent suspect and foils are equivalent such that they are chosen with equal probability. This, of course, is the basis for defining a fair lineup. We will use the terms designated-innocent and estimated-innocent to distinguish between these two kinds of lineups.

It is crucial, in evaluating the outcomes of eyewitness identification experiments, to examine the entire pattern of response probabilities, rather than any particular response probability. Consider, for example, an experiment with conditions A and B, where the correct identification rate is .30 for condition A and .50 for condition B. One might conclude that performance in Condition B is more accurate than for Condition A. But this conclusion would be undermined if the foil identification rate increased from .05 in Condition A to .40 in Condition B. Such a pattern would suggest a change in decision processes, rather than any improvement in identification performance. Similar ambiguities arise when any single response probability is considered in isolation. Indeed, many of the current controversies in eyewitness identification are concerned with increases in accuracy versus changes in response criterion (see Clare & Lewandowsky, 2004; Clark, 2005; Ebbesen & Konecni, 1996; Meissner, Tredoux, Parker, & MacLin, 2005). Resolution of these controversies will be facilitated by considering the complete patterns of all responses, rather than any single response.

The analyses to follow examine patterns of responses in two ways: First by looking at patterns of covariation across studies, and second by looking specifically at the diagnosticity or probative value of various eyewitness identification decisions. As noted by Wells and Lindsay (1980) and Wells and Olson (2002), the diagnosticity of witness responses is of crucial importance in the criminal justice system, as it addresses the question: Given that the witness has made response R , what is the likelihood that the suspect (or defendant) is guilty?

Our purpose in these analyses is not to test any particular effect. Nonetheless, because eyewitness identification procedures are defined in large part by two questions—regarding who is in the lineup and how it is administered—we focus on these two aspects of the task throughout our correlational and diagnosticity analyses. We consider lineup procedures through

the comparison of simultaneous versus sequential lineups, and we consider lineup composition by comparing studies that designated an innocent suspect versus those that did not. We will not at this point embark on long reviews of the literature regarding lineup procedure and composition, but will discuss each only briefly.

Simultaneous lineups present all lineup members at the same time and typically require one decision after the witness has considered all of the lineup members. In a sequential lineup, the lineup members are presented one-at-a-time, and the witness makes an identification decision for each lineup member as each is presented. Data from a number of studies has shown lower rates of false identification of innocent suspects using sequential lineups (see Steblay et al., 2001, for a meta-analysis), and consequently some jurisdictions have changed, or are currently proposing to change, from the use of simultaneous to sequential lineups.

The analyses of lineup composition are relevant to real crime investigations as well as the experimental, laboratory simulations of those investigations. As will be shown, lineups with designated innocent suspects have higher false identification rates and appear more biased than studies without designated innocent suspects, which are unbiased by the assumptions of the 1/K estimation procedure. The designated-estimated distinction can therefore serve as a proxy for lineup composition bias. Also relevant to the lineup composition question is how the foils are selected, and we will toward the end of the paper discuss how various procedures affect patterns of diagnosticity.

Method

Inclusion and exclusion criteria

The analyses are based on 94 experiments from 49 published studies. Only experiments that presented both TP and TA lineups were included in the analysis. In addition we considered only experiments using single-suspect lineups with adults as subjects. We excluded from the analysis studies which presented incomplete data, or presented data only in aggregate form, collapsing over conditions which produced significantly different results.

Articles were located by searches of Psych Info, as well as perusing the relevant journals (for example *Law and Human Behavior*, *Journal of Applied Psychology*, *Journal of Experimental Psychology: Applied*, *Applied Cognitive Psychology*). We consider only published studies in light of the clear position taken by the U.S. Supreme Court in *Daubert v. Merrel-Dow Pharmaceuticals*, et al. (1993) in which the Court established peer review as a key consideration in evaluating the reliability of scientific evidence in decisions to admit expert testimony. In *Daubert*, the Court deemed a reanalysis of epidemiological studies to be unreliable and therefor inadmissible in large part because the reanalysis had not been published. Clark (2005) recently showed how a single unpublished study of questionable reliability can dramatically change the conclusions of a meta-analysis.

Of the 94 experiments, 81 presented simultaneous lineups and 9 presented sequential lineups.¹ Also, 56 experiments reported identification rates for a designated innocent suspect, whereas

¹It is clear that the numbers of simultaneous and sequential lineups do not add to the total of 94. The reason is that we separated out a set of lineups that were presented in a manner that might be described as a hybrid of simultaneous and sequential procedures (Yarmey et al., 1996). Witnesses were presented with the photographs sequentially, but at the end of the sequence were allowed to go back through the photographs again. The authors noted that the use of this hybrid procedure was motivated by Canadian law at that time. The means for the hybrid procedure were as follows: In TP lineups, suspect, foil, and nonidentifications were: .390, .390, and .220, and in TA lineups: .193, .493, and .315.

38 did not designate an innocent suspect, requiring that innocent suspect identification rates be estimated, which we did using the 1/K procedure.

Random- versus fixed-effects analyses

Most of the analyses presented here were based on a random-effects, as opposed to a fixed-effects model. The two models differ in (a) the unit of analysis, (b) the proportional contribution of each data set, (c) the statistical power, and (d) the breadth of generalization of the statistical conclusions.

In a random-effects analysis, the unit of analysis is the study, rather than the participant within the study, as would be the case in a fixed-effects analysis. Typically, in a fixed effects analysis, the effect size estimates for each study are weighted by some function of the sample size. It comes as no surprise that random-effects analyses are more conservative, and considerably less powerful than fixed-effects analyses.

Random effects analyses differ from fixed-effects analyses in two other respects. For the random effects analysis each study contributes equally to the overall effect size, whereas for a fixed-effects analysis, studies with larger numbers of participants will contribute proportionally more than studies with smaller numbers of participants. The two statistical models differ also in the breadth of generalization. The statistical conclusions of random effects analyses generalize to the population of possible studies, whereas the conclusions of fixed-effects analyses generalize more specifically to the population of participants that could have been in the studies that were included in the meta-analysis. We balanced this trade-off by using random effects analyses whenever possible, when the set of studies was sufficiently large to allow it. However, for some analyses, the critical set of studies was quite small, such that random-effects analyses would be uninformative. For those analyses, fixed-effects analyses were used, trading off broader generalization for increased statistical power.

Outline of analyses

Following a brief overview of the response probabilities and the characteristics of the response distributions, the main analyses are divided into two sections. Section I examines patterns of covariability in response probabilities as a means of separating out the contributions of memory and decision processes. Section II expands upon earlier work by Wells and Lindsay (1980) and more recently by Wells and Olson (2002) to examine the information value or diagnosticity of different responses.

Results

The weighted and unweighted means and medians were calculated for each of the four response probabilities. There were only very small differences between the means and medians, and between weighted and unweighted statistics, so only the unweighted means are reported in the leftmost columns of Table 2. These nondifferences between means and medians are consistent with skew statistics showing that, with a few exceptions, none of the response distributions was particularly skewed or otherwise distorted by outlier probabilities. The exception to the no-skew results were findings of a slight positive skew in the don't know distributions (Skew/TP = 1.570; Skew/TA = 1.310), which were based on far fewer observations ($n = 13$) than other responses, and the suspect identifications in TA lineups (Skew = 2.267). The latter result arises primarily from aggregating over estimated and designated innocent suspect identifications. The

Table 2 Means for response probabilities, h and likelihood ratios and conditional probabilities for target-present and target-absent comparisons

	All Lineups ($n = 94$)							
	TP	TA	h	t	LR	t	CP	t
Suspect	.461	.134	.783	15.543 ^a	5.427	8.791 ^a	.773	19.239 ^a
Foil	.212	.345	.331	8.763 ^a	1.962	6.758 ^a	.632	8.202 ^a
Don't know ^b	.171	.192	.103	1.098	1.635	1.424	.575	1.298
No ID	.327	.520	.440	10.120 ^a	2.100	4.422 ^a	.618	8.522 ^a

Note. LR (likelihood ratio) = TP/TA for suspect identifications, and TA/TP for all other responses. CP (Conditional probability) = TP/(TP + TA) for suspect identifications and TA/(TP + TA) for all other responses.

^a $p < .001$.

^b Means based on data from 13 cases.

nondifferences between weighted and unweighted means, suggest that at least when considering the entire corpus, studies with large numbers of subjects did not produce results that were systematically different from studies with small numbers of subjects.

Correlations in response probabilities

Eyewitness identification is both a memory and a decision task, and is therefore sometimes described within the framework of Signal Detection Theory (Green & Swets, 1966) which is useful for understanding the separate and combined contributions of memory and decision processes (see for example Brown, Deffenbacher, & Sturgill, 1977; Clare & Lewandowsky, 2004; Clark, 2003, 2005; Malpass & Devine, 1981b). For example, an increase in correct identification rates can arise either because memory is more accurate or because witnesses are more willing to pick (see Clark, 2005). These two components of the eyewitness identification task are investigated meta-analytically by examining the correlations of response probabilities. Of course, response rates within a lineup condition must be negatively correlated, for quite uninteresting reasons. For example in a TP lineup, if the suspect identification rate is very high, say .90, the rate of false nonidentifications must be very low, and can be no higher than .10. Consequently, the interesting, and potentially illuminating, correlations are between TP and TA lineups. We note at the outset of this section that none of the correlational results varied across simultaneous and sequential lineups, or across designated versus estimated TA lineups. Thus, the results are presented only for the entire set of 94 TP-TA comparisons.

The analyses to follow are driven by the following simple predictions: To the extent that witness decisions are a function of the accuracy of the underlying memory trace, *correct* responses (i.e., suspTP, noidTA) should be correlated across TP and TA lineups. By contrast, to the extent that witness decisions are driven by particular response biases or decision strategies, response probabilities for the *same* response (i.e., suspTP, suspTA, etc.) should be positively correlated across TP and TA lineups.

Regarding the first half of the prediction, there are two ways in which a witness may make a correct identification response: either by identifying the suspect in a target-present lineup, (suspTP) or by making a correct nonidentification for a target-absent lineup (noidTA). It is a reasonable hypothesis that these two types of correct responses are correlated. Such a correlation would be consistent with the mirror effect pattern typically shown in recognition memory experiments (Glanzer & Adams, 1985, 1990). The mirror effect pattern shows that if overall performance, typically measured by d' , is better for condition A than for condition B, the

condition A advantage is typically produced by both increases in hit rate and decreases in false alarm rate. Thus, one might expect that conditions that would allow witnesses to accurately identify the target when present would also allow witnesses to reject lineups in which the target was not present. It is not difficult to imagine what such a condition might be: the accurate representation of information in the memory trace. As the accuracy of the memory trace increases, correct identifications and correct nonidentifications should both increase.

However, this turned out not to be the case, as the correlation between correct identification rates and correct nonidentification rates, considering all 94 experiments, was essentially zero ($r(92) = -.053, p = .614$). At first, this result seems surprising, but a moment's reflection provides a straightforward explanation of the lack of relationship. A correct identification requires that the witness picks *someone*, whereas a correct rejection requires that the witness picks *no one*. Thus, factors that affect criterion placement, i.e., the willingness to make any identification at all, would affect these two response probabilities in opposite directions, producing a negative correlation. The near-zero correlations observed may be the combined result of variation in the conditions that would affect the accuracy of memory and variation in conditions that would affect the decision criterion. Variation in memory factors would contribute to a positive correlation, whereas factors affecting decision criterion would contribute to a negative correlation.

This suggests that a measure of accuracy in TP lineups that is less affected by the placement of a criterion should be positively correlated with correct rejections in TA lineups. One such measure of accuracy is the conditional probability of selecting the target from the TP lineup, given that the witness selected anyone, $\text{suspTP}/(\text{suspTP} + \text{foilTP})$. As predicted, this conditional probability of correct identification was positively correlated with the correct nonidentification rate for the entire sample ($r(92) = .257, p = .012$).

The contributions of memory accuracy versus response criterion can be further examined through the TP-TA correlations for each response type. The correlations for each response category—suspect identifications, foil identifications, don't know, and nonidentifications—were computed across TP and TA lineups, and are shown in Table 3.

Consider the most straightforward case, which is for nonidentifications. To the extent that nonidentifications are based on the accuracy of the witness's memory, these responses should be negatively correlated. As memory accuracy increases, the correct nonidentifications should increase (noidTA) and the false rejection rates (noidTP) should decrease. Quite to the contrary, however, TP and TA nonidentification rates were highly correlated, $r(92) = .466, p < .001$. This result is consistent with the idea of a shifting response criterion and is inconsistent with what one would expect based on variation in the accuracy of witnesses' memories.

Table 3 Correlations in response probabilities for target-present and target-absent lineups

	All data ($n = 94$)	Sim ($n = 81$)	Seq ($n = 9$)
Suspect	.173 ^a	.178	.074
Foil	.559 ^d	.499 ^d	.780 ^b
Don't know ^e	.750 ^c	.750 ^c	–
No ID	.466 ^d	.442 ^d	.306

^a $p < .10$.

^b $p < .05$.

^c $p < .005$.

^d $p < .001$.

^eDon't know correlations based on 13 cases.

Foil identifications were also highly correlated ($r(92) = .559, p < .001$). The high correlation suggests that foil identifications provide a measure of witnesses' willingness to make an identification, irrespective of the presence or absence of the target. Likewise, the correlation between TP and TA don't know responses was also quite strong ($r(11) = .750, p = .003$).

The only responses that did not show a TP-TA correlation at the .05 level were the suspect identifications, $r(92) = .173, p = .096$. Presumably, the lack of correlation is once again the product of opposing factors, a shifting response criterion—that would produce a positive correlation, and the accuracy of the memory trace—that would produce a negative correlation.

Diagnosticity of eyewitness identification responses

Independently of whether responses are correlated across TP and TA lineups, a given response may occur much more often in one lineup condition than in another. To the extent that this is the case, that response is diagnostic with respect to the suspect's guilt or innocence.

Wells and Lindsay (1980) developed two measures of diagnosticity, one based on ratios and one based on conditional probabilities. Diagnosticity ratios are given by the ratio of response probabilities for TP and TA lineups, typically expressed with the larger divided by the smaller probability, as shown in Wells and Lindsay's Equation 6 (p. 779). The conditional probability may be viewed as a posterior probability of the suspect's guilt given that the witness has provided a particular response to the lineup.

This section of the meta-analysis may be viewed as an extension of previous work by Wells and Lindsay (1980) and Wells and Olson (2002) who considered the diagnosticity for suspect, foil, don't know, and nonidentification responses. These previous analyses showed all four response types to be diagnostic with respect to the presence of the target. Specifically, suspects were identified more often in TP lineups, and the other three response types occurred more often in TA lineups.

We build upon the previous work in three ways: (1) First, we considered three measures of diagnosticity, using Cohen's h , conditional probabilities, and likelihood ratios, (2) we considered a much larger number of studies, and (3) we compared each of the various response types in order to draw conclusions as to the relative diagnosticity of each response type. We also followed Wells and Olson by examining whether diagnosticity varied across simultaneous and sequential lineups, and across variations in lineup composition.

Measures of diagnosticity: Likelihood ratios, conditional probabilities, and Cohen's h

Likelihood ratios are commonly reported as measures of response diagnosticity in eyewitness identification experiments. They have many useful properties, the most obvious of which is their straightforward interpretation. For example, with suspect identification rates of say .45 and .15 for TP and TA lineups, one can assert that witnesses were three times more likely to identify the suspect when he was guilty than when he was innocent.

There are, however, some potential problems with likelihood ratios. First, they are unstable when considering responses that occur with low probability. Second, for ratios less than 1, large differences are compressed, producing distributions that are very positively skewed. These properties can be troublesome if one computes a summary statistic, i.e., the mean of the diagnosticity ratios across several experiments. Ratios less than 1 will be buried by ratios greater

than 1 and the means will be considerably larger than the medians, suggesting that they are rather unrepresentative summary statistics.

Because of these problems we also calculated conditional probabilities (CP). These conditional probabilities also have a straightforward and forensically relevant interpretation. In the case of suspect identifications, the conditional probability answers the question: Given that the witness identified the suspect, what is the probability that he is guilty? Alternatively, for suspect identifications, the conditional probability can be inverted to give a measure of the *innocence risk*, i.e., the likelihood that the suspect is innocent, given that he was identified.

We also calculated Cohen's h (Cohen, 1988), which is the difference between the two arcsin transformed probabilities. Of the various effect-size statistics that could be calculated, we picked h for three reasons: (1) it is frequently used in meta-analyses of proportions, (2) Z scores can be easily and directly calculated from h (see Cohen, 1988, p. 209), and (3) it expands differences at the high and low ends of the probability scale, reflecting the fact that a .05 difference between .05 and .10 seems bigger than the .05 difference between .45 and .50.

For all of the diagnosticity analyses to follow, the measures of diagnosticity were calculated consistent with the view (see Wells & Lindsay, 1980; Wells & Olson, 2002) that suspect identifications are diagnostic of guilt, whereas all other responses are diagnostic of innocence. Thus, h for suspect diagnosticity was calculated by subtracting TA from TP, but all other h were calculated by subtracting TP from TA, in order to produce effect sizes with positive signs. Similarly, likelihood ratios (LR) were calculated as TP/TA for suspect identifications, but TA/TP for all other responses, producing likelihood ratios greater than 1; and conditional probabilities (CP) were calculated as TP/TP + TA for suspect identifications and TA/TP + TA for all other responses, producing conditional probabilities generally greater than .5

Overall diagnosticity results

The analyses were based on 94 separate TP-TA comparisons, with the exception of the don't know responses, which were based on only 13. The results are shown in Table 2. The dependent measures, Cohen's h , likelihood ratios, and conditional probabilities, lead to the same conclusions, and therefore the conclusions can be viewed as quite general, rather than peculiar to any particular way of computing diagnosticity. This is not to say, however, that the three measures are interchangeable. As expected, the distributions for likelihood ratios were very positively skewed such that means were 20 to 40 percent larger than medians. In contrast, means and medians were nearly identical for h and for conditional probabilities. Also, although all the response measures were correlated, the correlations were higher between h and CP (.866 to .919) than between LR and CP (.638 to .885) or between LR and h (.570 to .785). These preliminary analyses have important implications, specifically that h and conditional probabilities may be used interchangeably as measures of diagnosticity, but that likelihood ratios should be used with caution. Given the problems in aggregating likelihood ratios, discussion of results throughout the paper will focus on h and conditional probabilities, although the likelihood ratio analyses are maintained in the tables.

The random-effects analyses in Table 2 show that, with the exception of don't know responses, each response was diagnostic, with all p 's < .001. Don't know responses were not diagnostic irrespective of which statistic was used. Don't know responses, of course, were based on far fewer comparisons than were the other response categories. Thus, to further investigate the diagnosticity of don't know responses, we conducted a fixed-effects analysis, converting the h effect sizes to z -scores (Cohen, 1988, p. 209), and computing a meta-analytic Z using the Stouffer method (Rosenthal, 1991; Stouffer, Suchman, DeVinney, Starr, & Williams, 1949). The average effect size was small, $h = .103$, and meta-analytic Z was nonsignificant, $Z = 1.098$, $p = .136$.

We also considered whether there was sufficient heterogeneity in the effect to suggest that the nonsignificant difference was the product of significant differences in opposite directions. A diffuse test for heterogeneity (see Rosenthal, 1991, p. 73, Eq. 4.14) did not reach statistical significance, $\chi^2(13) = 18.700, p = .133$, suggesting that this was not the case.

Relative diagnosticity

We compared the measures of diagnosticity to each other, suspect vs. foil, suspect vs. non-identification, and foil vs. nonidentification. In these analyses, the null to be tested is that the average difference between the response outcomes within the measures of diagnosticities (e.g., Cohen's h for suspect—Cohen's h for foil) will be zero, evaluated by single-sample t -tests. This analysis showed that suspect identifications were far more diagnostic than any other response. Specifically, the observed average mean difference between suspect Cohen's h and foils Cohen's h was statistically significant, ($t(93) = 10.113, p < .0001$), as was the average mean difference between suspect Cohen's h and nonidentification Cohen's h ($t(93) = 9.166, p < .0001$). Further, the average difference between foil Cohen's h and nonidentification Cohen's h did not reach statistical significance ($t(93) = 1.770, p = .08$).

Diagnosticity in simultaneous and sequential lineups

We turn next to an analysis of the diagnosticity of suspect, foil, and nonidentification responses in simultaneous and sequential lineups. The relevant data are in Table 4. For completeness, the likelihood ratios are reported in the Table; however, discussion of the results will focus on Cohen's h and the conditional probabilities.

The conditional probabilities and h show virtually no difference in the diagnosticity of suspect identifications, greater diagnosticity of foil identifications in simultaneous lineups than in sequential lineups, and greater diagnosticity of nonidentifications in sequential lineups than in simultaneous lineups. To evaluate these results statistically, we computed a Z score from each h comparing TP and TA lineups, calculated the Z from h (following Cohen, 1988, Eq. 6.5.3, p. 209), and then computed the contrast meta-analytic Z (following Rosenthal, 1991, Eq. 4.26, p. 79). This analysis was consistent with the patterns for conditional probabilities and h shown in the table (with positive Z indicating higher diagnosticity for sequential lineups, and

Table 4 Diagnosticity in simultaneous and sequential lineups

	TP	TA	h	t	LR	t	CP	t
Simultaneous ($n = 81$)								
Suspect	.471	.137	.801	14.544 ^b	5.473	8.469 ^b	.775	17.749 ^d
Foil	.208	.350	.351	8.335 ^b	1.961	7.221 ^b	.639	7.818 ^b
No ID	.320	.513	.435	9.448 ^b	2.100	3.908 ^b	.617	7.635 ^b
Sequential ($n = 9$)								
Suspect	.396	.081	.773	4.756 ^b	6.433	2.450	.794	6.826 ^b
Foil	.166	.242	.201	2.091	2.263	1.448	.599	2.038 ^c
No ID	.431	.669	.592	3.259 ^b	2.359	2.358	.644	3.182 ^a

^a $p < .10$.

^b $p < .05$.

^c $p < .01$.

^d $p < .001$.

one-tailed p values). Specifically, suspect identifications showed no diagnosticity difference in simultaneous and sequential lineups ($Z = .391, p = .348$), foil identifications were more only slightly more diagnostic in simultaneous than in sequential lineups ($Z = 1.319, p = .094$), and nonidentifications were more diagnostic in sequential than in simultaneous lineups ($Z = 1.917, p = .028$).

This pattern of results is inconsistent with a previous meta-analysis by Steblay et al. (2001) and an analysis by Wells and Olson (2002). There are, of course, a number of differences between the analyses summarized above and those of Steblay et al. and Wells and Olson. One of the primary differences is that both of the previous analyses were based only on data from experiments that directly compared simultaneous and sequential lineups, whereas the results shown in Table 4 considered results from simultaneous and sequential lineups irrespective of whether they were part of a direct comparison.

Because the use of sequential lineups is an important component in the reform of police procedures and policies in many state and local jurisdictions (i.e., California, New Jersey, North Carolina, Virginia, Wisconsin), these diagnosticity comparisons are worthy of closer examination. Because of the small number of studies, the following analyses are based on a fixed- rather than random-effects analysis. We first calculate and present the results of a direct-comparison meta-analysis of diagnosticities in simultaneous and sequential lineups, and then turn to how these analyses compare with studies by Steblay et al. and Wells and Olson. Jumping ahead, the comparisons across the various analyses show considerable variation. For clarity, the conclusions of the analyses are summarized in Table 6.

Direct simultaneous-sequential lineup comparisons

Direct simultaneous-sequential comparisons were made in eight studies that met our inclusion criteria, and the mean response probabilities are shown in Table 5. The relative diagnosticities of each response type for simultaneous and sequential lineups were analyzed by first computing the conditional probabilities, [TP/(TP + TA) for suspect identifications and TA/(TA + TP) for foil and nonidentifications], and then calculating Cohen’s h for the difference between those two conditional probabilities. A meta-analytic Z score was then computed from Cohen’s h . The means and medians for Cohen’s h , and the meta-analytic Z scores are given in Table 5. Again, positive h and Z indicate greater diagnosticity for sequential than for simultaneous lineups.

The diagnosticity for suspect identifications was higher for sequential lineups than for simultaneous lineups ($h = .230, Z = 2.242, p = .012$, one-tailed). Seven of the eight studies showed this result. No differences in diagnosticity were found for either foil identifications or nonidentifications ($h = -.035, Z = .329$ for foil identifications and $h = .206, Z = .549$ for nonidentifications). Thus, these analyses indicate greater diagnosticity for sequential lineups

Table 5 Response probabilities and conditional probabilities for TP and TA simultaneous and sequential lineups in direct-comparison studies

	Sim			Seq			h_M	h_{Mdn}	Z
	TP	TA	CP	TP	TA	CP			
Suspect	.496	.217	.700	.391	.081	.789	.230	.314	2.242
Foil	.248	.399	.632	.123	.199	.607	-.035	-.129	.329
No ID	.256	.384	.535	.478	.713	.600	.206	.007	.549

Note. TP = target-present, TA = target-absent, CP = conditional probability: TP/(TP + TA) for suspect identifications, and TA/(TP + TA) for foil and nonidentifications; h_M and h_{Mdn} denote mean and median effect sizes (h).

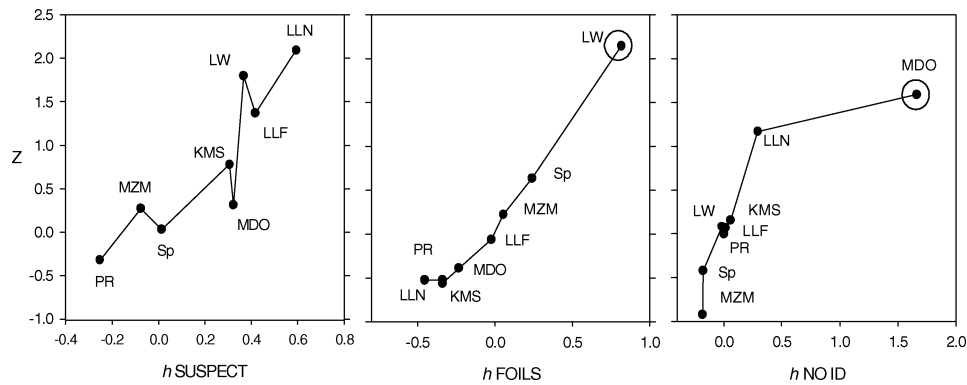


Fig. 1 $h \times Z$ plot for suspect, foil, and nonidentification response diagnosticity for direct comparison of simultaneous and sequential lineups. Each study is identified by a point in h, Z space, with h on the x -axis and Z on the y -axis. Studies are identified as KMS (Kneller, Memon, & Stevenage, 2001), LLF (Lindsay, Lea, & Fulford, 1991), LLN (Lindsay et al., 1991), LW (Lindsay & Wells, 1985), MDO (Melara et al., 1989), MZM (MacLin et al., 2005), and Sp (Sporer, 1993)

than for simultaneous lineups, but only for suspect identifications. Before addressing how these results compare with others we first note three additional aspects of the present results for what they reveal about the corpus of simultaneous-sequential studies. These points will be important later when comparing the present meta-analysis to the analysis reported by Wells and Olson.

First, the mean for Cohen's h for nonidentifications was only slightly lower (.206) than for suspect identifications (.230), but the Z scores were quite different. Second, for both foil and nonidentifications, the mean and median for h were quite different. Third, the Z for foil identifications was positive (although nonsignificant), even though the mean and median for h were negative. What might be behind these peculiarities in the results? Discrepancies between means and medians suggest the presence of outlier scores which may render the mean a less-accurate summary statistic than the median. Discrepancies between effect sizes and Z scores can arise if there is an association between effects and sample sizes, such that large-sample studies show different results than small-sample studies.

Effect size outliers were identified by calculating a modified Z score which uses the median rather than the mean as its measure of central tendency (Iglewicz & Hoaglin, 1993).² This analysis revealed no outliers for suspect identifications. However, the effect size for foil identifications for Lindsay and Wells (1985), $h = .816$ and the effect size for nonidentifications for Melara, DeWitt-Rickards, and O'Brien (1989), $h = 1.662$, were both shown to be outliers in their respective distributions. Our purpose in identifying these outliers is neither to exclude them nor to question the validity of the results, but rather to show they contribute to the aggregate of data. A graphic illustration of the distributions on h and Z , with the outliers identified, is shown in Fig. 1. The figure plots diagnosticity for each study, for suspect, foil, and nonidentification responses as points in h, Z space, with h on the x -axis and Z on the y -axis. The figure serves the same purpose as a stem-and-leaf plot, but with the added information about how effect sizes are translated into

²The modified Z is like the standard Z , but uses the deviations about the median (rather than the mean) as in the denominator. The advantage is that the median, as a measure of central tendency, is less influenced than the mean by the presence of the extreme score that one is trying to identify. The modified Z scores for the two scores we have identified as outliers were 3.04 (Lindsay & Wells, 1985) and 7.28 (Melara et al., 1989). The Lindsay and Wells result is lower than the 3.5 cutoff used by Iglewicz and Hoaglin (1993); however its large sample size gave it considerable weight, particularly in the Wells and Olson (2002) calculations.

Z scores. The Figure shows the Lindsay and Wells foil identification result to be an outlier not only for h , but also for Z , whereas the Melara et al. result is shown to be an outlier on h , but much less of an outlier on Z . The reason is that the Lindsay and Wells results were based on a large number of participants (60 per condition, 240 total) whereas Melara et al.'s results were based on a very small number of participants (8 per condition, 32 total). Thus, the Lindsay and Wells results contribute heavily to the average h and Z , whereas the Melara et al. results contribute only to h , but not to Z . The positive Z combined with negative h for foil identifications was due to the large contribution made by the Lindsay and Wells study. For nonidentification responses the moderate effect size (nearly equal to that for suspect identifications) was due to the Melara et al. study, which contributed little to the meta-analytic Z due to its very small sample size. As we will show, these two studies played a very large role in the Wells and Olson's analyses.

Comparison to Wells and Olson's (2002) analysis

The results of our direct comparison differ from the results reported by Wells and Olson (2002) primarily in that they reported greater diagnosticity for foil identifications in sequential lineups than in simultaneous lineups, whereas the present meta-analysis showed no difference. Their conclusions are summarized in Table 6.

This difference may have been due to the five studies added to the present analysis, or to differences in how the results were aggregated and analyzed. We reanalyzed the same three studies from the Wells and Olson analysis, computing Cohen's h and meta-analytic Z and obtained results very similar to those shown in Table 5. (For comparison, for the three studies in the Wells and Olson analysis, for suspect identifications: $h = .368$, $Z = 2.011$, $p = .022$, one-tailed; for foil identifications: $h = .186$, $Z = .975$, $p = .165$, one-tailed; and for nonidentifications: $h = .553$, $Z = .909$, $p = .187$, one-tailed). Thus, our reanalysis of their three studies leads to the same statistical conclusions as our analysis based on all eight studies, including a noneffect for the diagnosticity of foil identifications. The question then is what analysis lead Wells and Olson to conclude that foil identifications were more diagnostic in sequential than in simultaneous lineups? The answer lies not in which studies were analyzed, but rather in how the studies were analyzed. We turn to this next.

To compute their conditional probabilities Wells and Olson summed responses across three studies (Lindsay & Wells, 1985; Lindsay et al., 1991; Melara et al., 1989). This aggregation of raw-score frequencies across contingency tables can produce paradoxical results, i.e., aggregate tables that look quite unlike any of the tables that contributed to the aggregate (see Hintzman, 1980, Rosenthal, 1991; Simpson, 1951; Yule, 1903). For that reason, meta-analytic aggregation of effect sizes is preferred over summing raw frequencies (Rosenthal, 1991).

The effect of raw-score summation in the Wells and Olson analysis was that studies with large numbers of subjects made substantially larger contributions to the aggregate table than studies with small numbers of subjects. The three studies considered by Wells and Olson had

Table 6 Summary of conclusions regarding diagnosticity of suspect, foil, and nonidentification responses in simultaneous and sequential lineups

	Present analysis		Stebly et al.	Wells and Olson
	Full Sample	Direct comparison		
Suspect	Seq = Sim	Seq > Sim	Seq > Sim	Seq > Sim
Foil	Seq = Sim	Seq = Sim	Seq < Sim	Seq > Sim
No ID	Seq > Sim	Seq = Sim	Seq = Sim	Seq = Sim

240, 180, and 32 subjects for Lindsay and Wells (1985), Lindsay et al. (1991) and Melara et al. (1989), respectively. Because the Lindsay and Wells results were outliers in the distribution and contributed over half of the data to the raw-score summation, they dominated the aggregate. The other two studies actually showed foil diagnosticity to be lower in sequential than in simultaneous lineups; thus Wells and Olson's conclusion of higher diagnosticity was due entirely to the Lindsay and Wells results.

Comparison with Steblay et al. (2001)

Steblay et al. conducted a meta-analysis for 28 data sets (including several unpublished studies). Although they did not calculate diagnosticity statistics, some patterns regarding diagnosticity can be observed in their Table 1 (p. 463), which shows the means for suspect, foil, and nonidentification response probabilities for TP and TA lineups, and for simultaneous and sequential lineups. Conditional probabilities can be computed from these means, with the caveat that conditional probabilities calculated from means are not the equivalent of computing means for conditional probabilities based on data from each study.

With that caveat, the most straightforward case in their analysis, one that does not require any estimation or additional assumptions, is for nonidentification responses, which were nearly the same for simultaneous (.653) and sequential (.610) lineups. Conditional probabilities for suspect identifications can be computed with one additional caveat—the guilty suspect identification rates in TP lineups are based on all TP comparisons, whereas innocent suspect identification rates are based only on those studies that designated an innocent suspect. With that caveat, the conditional probabilities were higher for sequential lineups (.795) than for simultaneous lineups (.649). Foil identifications can only be estimated, and our estimation was made by multiplying the identification rates in TA lineups by a constant (5/6) to estimate response proportions based on a fair six-person lineup. With those assumptions, foil identifications appeared to be more diagnostic in simultaneous (.680) than in sequential (.596) lineups. These conclusions are summarized in Table 6.

Summary of direct-comparison analyses

The three direct-comparisons between simultaneous and sequential lineups were consistent in that all three showed greater diagnosticity of suspect identifications for sequential lineups, and all three showed no difference between simultaneous and sequential lineups for the diagnosticity of nonidentification responses. The three studies showed different patterns concerning the diagnosticity of foil identifications: The present analysis showed no difference, estimates from Steblay et al. (2001) suggest greater diagnosticity in simultaneous lineups and Wells and Olson reported greater diagnosticity in sequential lineups. A detailed analysis showed that the Wells and Olson results were due primarily to one study, with unusual results (Lindsay & Wells, 1985) that dominated their analysis. The pattern of results based on the Steblay et al. analysis required a number of assumptions and was largely *our* analysis based on their means. Consequently, among the direct-comparisons, the present analysis appears to give the clearest picture regarding the diagnosticity of foil identifications in simultaneous and sequential lineups. Nonetheless, given the small number of comparisons, this issue would be better-informed with additional experimentation.

Table 7 Results for simultaneous lineups in studies that designated an innocent suspect, and either did or did not make a direct simultaneous-sequential comparison

	Sequential comparison		No sequential comparison	
	Target present	Target absent	Target present	Target absent
Suspect	.568	.272	.482	.165
Foil	.139	.257	.189	.308
No ID	.294	.472	.330	.526
S/(S + F)	.802	.529	.698	.328

Note. S/(S + F) is the suspect identification rate divided by the sum of the suspect plus foil identification rate, and is the conditional probability of identifying the suspect, given that any identification was made.

Differences relative to the full-sample

We are left here with inconsistent results for the direct-comparison versus the full-sample analyses. Specifically, in contrast to all of the direct comparisons, the full-sample analysis showed no simultaneous-sequential difference in the diagnosticity of suspect identifications, no difference for foil identifications, and greater diagnosticity of nonidentifications for sequential lineups. What might account for these differences? The discussion to follow is focused on the differences for suspect identifications.

These differences may be due in large part to how the innocent suspect and foils were selected in target-absent lineups. As will be shown later, the conditional probability of a suspect identification given any identification is higher when the innocent suspect is designated than when the innocent suspect identification rate is estimated. Differences in results could be obtained if the proportion of studies requiring the 1/K estimation were higher in one sample than in another. In fact, 57 percent of simultaneous lineup studies designated an innocent suspect, whereas 67 percent of sequential lineup studies designated an innocent suspect. Although this slight imbalance likely contributed somewhat to the difference between the whole-sample versus the focused-sample results, subsequent analyses presented next suggest additional differences due to lineup composition.

We considered in both the full set of studies, and those which made direct simultaneous-sequential comparisons, those which also designated an innocent suspect. The critical results are for the conditional probabilities of identifying the suspect given any identification, $Susp/(Susp + Foil)$. This conditional probability is a measure of lineup composition, specifically the relative similarity of the suspect and foils. The relevant probabilities are shown in Table 7. The critical comparison is between simultaneous lineups that were part of a simultaneous-sequential comparison versus simultaneous lineups that were not part of a simultaneous-sequential comparison. The results are straightforward. First, these two sets of simultaneous lineups did not differ for the target-present condition ($t(44) = 1.270, p = .211$). However, the conditional probability of identifying the innocent suspect ($Susp/(Susp + Foil)$) in TA lineups was significantly higher for simultaneous lineups that were part of a simultaneous-sequential comparison than for simultaneous lineups that were not part of a simultaneous-sequential comparison ($t(44) = 2.118, p = .040$).³ This result suggests that studies that made a simultaneous-sequential comparison, and also designated an innocent suspect used TA lineups that were more

³Simultaneous lineups in simultaneous-sequential comparisons were on average larger (6.8) than simultaneous lineups in studies without a sequential comparison (6.1), which would tend to reduce $Susp/Susp + Foil$ for those

biased against the innocent suspect than did comparable studies that did not make a direct simultaneous-sequential comparison. This suggests that the diagnosticity advantage for suspect identifications for sequential lineups shown in the direct-comparison meta-analysis may be due in part to the similarity relations in those lineups. Specifically, the result may be restricted to only those cases in which the TA lineups are more biased against the innocent suspect. When the TA lineups are less biased, the sequential advantage in the diagnosticity of suspect identifications may not hold. We should note one last important caveat. The results of the analysis shown in Table 7 suggests that the full-sample simultaneous-sequential comparison (shown in Table 4) makes a comparison between nonequivalent cases, as the innocent suspects appear to have higher similarity to the target in sequential lineups than in the full set of simultaneous lineups.

Diagnosticity and lineup composition

Lineup composition and similarity relationships in TP and TA lineups are determined in large part by the similarity of the innocent suspect to the actual perpetrator and by how the foils are selected. In this section we examine both of these relationships and how they affect the diagnosticity of witness responses.

Diagnosticity in designated- versus estimated-innocent lineups

Although the issue has not been studied directly, it seems almost certainly the case that false identifications of innocent suspects will increase the more similar the innocent person is to the perpetrator. We examine this relationship by comparing studies that designated an innocent suspect to those that did not. For the estimated-innocent lineups the false identification rate was estimated by procedures that assumed the lineups to be fair and unbiased with respect to the innocent suspect. As will be shown, the designated-innocent TA lineups were quite clearly biased with respect to the innocent suspect.

Our first question is whether the results of studies which required the 1/K estimation procedure differed from the results of studies which designated an innocent suspect, and provided suspect and foil identification rates directly. The answer to this question is shown in Table 8. To obtain the most direct comparison between designated-innocent and estimated-innocent experiments, the analyses to follow were restricted to six-person, simultaneous lineups. Unrestricted analyses of the full data set showed the same results.

The patterns of responses for target-present lineups did not differ between studies which designated an innocent suspect and those which required the fair-lineup estimation. There were no differences for suspect, foil, or nonidentifications ($t(61) = 1.010, 1.328, \text{ and } 0.098$, respectively). The conditional probability of identifying the suspect in TP lineups, $\text{SuspTP}/\text{SuspTP} + \text{FoilTP}$, was also the same irrespective of how the target-absent lineups were created ($t(61) = .969$). These nondifferences are important because the designation of the innocent suspect should have no effect on the results of target-present lineups. Any other result would have been curious.

In contrast, for TA lineups suspect identifications were higher ($t(61) = 3.339, p = .001$) and foil identification rates lower ($t(61) = 2.699, p = .009$) in studies which designated an innocent suspect, relative to those which required 1/K estimation. Rates of nonidentifications did not differ ($t(61) = 0.181$). Because the conditional probability of a suspect identification given

studies (in opposition to the obtained results) An analysis only for studies where $k = 6$ showed the same pattern of results, but with only three cases for the simultaneous-with-sequential condition.

Table 8 Identification data, Cohen's *h*, likelihood ratios, and conditional probabilities for studies with and without a designated innocent suspect

	TP	TA	<i>h</i>	<i>t</i>	LR	<i>t</i>	CP	<i>t</i>
Lineups with designated innocent suspect (<i>n</i> = 36)								
Suspect	.492	.190	.717	7.689	4.239	4.511	.730	8.132
Foil	.192	.314	.296	5.071	1.917	4.738	.623	4.929
No ID	.316	.496	.420	7.311	1.833	5.693	.618	6.839
Lineups without designated innocent suspect (<i>n</i> = 27)								
Suspect	.443	.085	.869	10.054	6.783	5.633	.816	16.396
Foil	.237	.428	.479	6.742	2.109	4.516	.679	5.809
No ID	.320	.487	.353	4.080	2.261	1.657 ^a	.583	2.637 ^b

Note. All tests significant $p < .001$, except.

^a $p = .109$.

^b $p = .014$.

any identification ($\text{SuspTA}/\text{SuspTA} + \text{FoilTA}$) is fixed at .167 for the 1/K estimated lineups, the conditional probability for designated-suspect lineups was evaluated by a single-sample *t*-test. Clearly, this conditional probability (.350) was much higher than the fair-lineup baseline ($t(35) = 4.918$, $p < .001$). Designated-innocent lineups were quite biased. What effect does this bias have on the diagnosticity of witness responses?

Diagnosticity measures based on Cohen's *h*, likelihood ratios, and conditional probabilities are shown in Table 8. As in all of the previous analyses suspect identifications were diagnostic of guilt and foil and nonidentifications were both diagnostic of innocence, in lineups that designated an innocent suspect and in those that did not. The diagnosticity differences between designated-innocent and estimated-innocent lineups, although inconsistent across measures of diagnosticity, showed a consistent pattern. The analysis using conditional probabilities showed lower diagnosticity for suspect identifications in designated-innocent lineups ($t(61) = 2.356$, $p = .022$), and the analysis using Cohen's *h* showed significantly lower diagnosticity of foil identifications in designated-innocent lineups than in estimated-innocent lineups, $t(61) = 2.008$, $p = .049$. Neither diagnosticity measure showed a difference between designated- and estimated-innocent lineups for nonidentification responses.

The results of designated-innocent lineups may have been obtained because experimenters, when designating an innocent suspect, often pick a person from their pool of stimuli, who, by similarity ratings or by intuition, is highly similar, or even the *most* similar, to the actual perpetrator. In other words, the higher false identification rates for the innocent suspect occur because experimenters make them occur. This might suggest that biased lineups occur only when experimenters design bias into their experiments. However, biased lineups may in fact be a natural outcome of the lineup construction procedures that police typically use. Here the issue is not about the relationship between the innocent suspect and the perpetrator, but rather about how the foils for lineups are selected.

Diagnosticity in suspect-matched and description-matched lineups

Police officers, in two surveys (Wogalter, Malpass, & McQuiston, 2004; Wogalter, Malpass, & Burger, 1993), reported that they select foils for lineups based on their similarity to the suspect. A number of research findings and theoretical analyses (Clark, 2003; Clark & Tunnicliff, 2001; Navon, 1992; Py, Demarchi, Ginet, & Wasiak, 2003; Wogalter, Marwitz, & Leonard, 1992),

Table 9 Basic identification results for suspect-matched lineups

	TP	TA	TP-TA	<i>h</i>	LR	CP
Suspect	.371	.115	.256	.642	5.833	.758
Foil	.324	.261	.063	-.145	0.849	.441
No ID	.303	.621	-.318	.662	2.218	.671
S/(S + F)	.534	.306				

Note. LR = TP/TA for suspect identifications, and TA/TP for foil and nonidentifications. CP (conditional probability) = TP/(TP + TA) for suspect identifications and TA/(TP + TA) for foil and nonidentifications.

suggest that biased lineups may be the norm rather than the exception, and moreover that biased lineups may be created not only when police procedures are violated, but even when they are followed *exactly*. Specifically, foil selection based on the similarity to the suspect can, by design, produce biased lineups in two ways.

First, it can produce lineups in which the suspect can be identified by nonwitnesses because the suspect represents the prototype or “parent” for the rest of the lineup (Wogalter et al., 1992). Second, consider the case in which the innocent person becomes a suspect because he fits the description of the perpetrator given by the witness. This innocent suspect is placed in a lineup with five fillers who look similar to that innocent suspect. Given this scenario, one may ask: How many people are in the lineup because they fit the description of the perpetrator? The answer is only one—the innocent suspect. *The other five people are in the lineup because they look similar to a person who fits the description of the perpetrator.* Given that the description of the perpetrator is a product of the witness’s memory, it follows that the person in the lineup who will be the closest match to the witness’s memory is the innocent suspect. It follows further that the innocent suspect is the person in the lineup who is the most likely to be identified.

Only six studies have created lineups by selecting foils based on their similarity to the suspect (Table 9), one of which (Lindsay, Martin, & Webber, 1994) was not included in the analysis because it did not report the relevant data, leaving only five studies for these analyses. Suspect-matched foil selection, although it may be common practice for law enforcement, produces lineups that are unlike the lineups in most eyewitness identification experiments. Because foils are selected based on their match to the perpetrator in TP lineups and based on their match to the innocent suspect in TA lineups, TP and TA lineups should consist of nonoverlapping sets of foils. By contrast in most experiments, TP and TA lineups share the very same set of foils.

The three measures of diagnosticity, Cohen’s *h*, likelihood ratios, and conditional probabilities, were calculated as before. Because the set of studies is very small, diagnosticity of responses was evaluated by fixed-effects analyses. Similar to previous fixed-effects analyses, we calculated *h* from the difference in conditional probabilities, calculated individual *Z* scores from *h*, and computed a meta-analytic *Z* by the Stouffer method.

The results showed the same pattern for suspect identifications and nonidentifications as shown in all of the previous analyses, with suspect identifications diagnostic of guilt ($Z = 7.303$, $p < .001$), and nonidentifications diagnostic of innocence ($Z = 7.849$, $p < .001$). However, contrary to previous analyses, foil identifications were higher for TP than for TA lineups, and were thus diagnostic of guilt ($Z = 1.945$, $p = .026$), rather than innocence as shown in all of the other analyses. This result should be considered with some caution because it is based on fixed-effects analysis of a small number of studies, and although three of the five studies showed the TP > TA pattern, the statistical significance of the pattern was due in large part

to the results of Clark and Tunnicliff (2001) who showed foil identification rates of .468 and .159 for TP and TA lineups. Even with those caveats, one point is clear: contrary to all other analyses, for suspect-matched lineups, foil identifications are not diagnostic of the suspect's innocence.

This result is important because although it may be unusual in the present set of studies it may be quite representative of lineups in actual police investigations. As noted earlier, the large majority of police officers surveyed indicated that they typically select foils based on their similarity to the photograph of the suspect.

As we noted earlier, suspect-matched lineups are also inherently biased against the innocent suspect. In addition they may also select a set of foils that is unnecessarily similar to the target in TP lineups, resulting in lower correct identification rates (see Luus & Wells, 1991, Wells, Rydell, & Seelau, 1993). In part for these reasons, many eyewitness identification researchers have recommended that foils be selected based on their match to a description of the perpetrator, rather than the appearance of the suspect (see Wells et al., 1998). Clark (2003) reviewed the studies which have directly compared suspect-matched and description-matched lineup composition, and we will not revisit the details of that review here. However, in the next section, we compare suspect-matched and description-matched lineups, again with primary focus on the diagnosticity of witness responses.

Only four of the five studies that directly compared suspect-matched and description-matched lineups meet our inclusion criteria. The Clark and Tunnicliff experiment considered only suspect-matched lineups, so is not included in this analysis. The means of the response probabilities are as follows: For suspect-matched lineups, $\text{suspTP} = .380$, $\text{foilTP} = .288$, $\text{noidTP} = .330$, $\text{suspTA} = .081$, $\text{foilTA} = .287$, and $\text{noidTA} = .630$. For description-matched lineups: $\text{suspTP} = .508$, $\text{foilTP} = .147$, $\text{noidTP} = .348$, $\text{suspTA} = .131$, $\text{foilTA} = .282$, and $\text{noidTA} = .585$. Because these results have been considered in detail elsewhere (Clark, 2003), we will focus only on the diagnosticity of witness responses. Again, with a very small set of studies we used a fixed-effects analysis to determine whether the diagnosticity of witness responses varied in the two kinds of lineup. The only clear difference was shown for foil identifications, which were diagnostic of innocence in description-matched lineups (conditional probability, $\text{foilTA}/\text{foilTA} + \text{foilTP} = .652$) but nondiagnostic in suspect-matched lineups (conditional probability = .488). Suspect-matched and description-matched lineups did not differ with respect to the diagnosticity of suspect ($Z = .665$) or nonidentification responses ($Z = .763$).

There may also be some bias problems associated with description-matched lineups (see Clark, 2003). For present purposes it suffices to note that the conditional probability of identifying the innocent suspect, $\text{SuspTA}/\text{SuspTA} + \text{FoilTA} = .317$, exceeds the fair lineup baselines which vary from .20 to .125 across the studies with lineup sizes from 5 to 8, and is somewhat higher than it is in suspect-matched lineups ($\text{SuspTA}/\text{SuspTA} + \text{FoilTA} = .220$).

General discussion

This paper began with a simple question: What do eyewitness identification experiments typically show? The results should—and do—vary considerably across experimental procedures. However, regularities in the patterns of witness responses persist across variations in procedures. Two sets of analyses were directed toward those regularities. The first set of analyses examined the covariability in response probabilities, and the second set examined response diagnosticity. We turn now to summarize the results and what they reveal about eyewitness identification.

Summary of results

Correlational analyses of witness responses

There was no correlation between correct identifications (in TP lineups) and correct nonidentifications (noidTA). We propose that this near-zero correlation is due to the combination of two factors acting in opposition: variation in the accuracy of the memory trace (which should produce a positive correlation) and variation in response criterion (which would produce a negative correlation). This separation of memory accuracy and response criterion is at the heart of signal detection theory, which has been applied widely, not only for recognition memory (Clark & Gronlund, 1996) and eyewitness identification (Brown, Deffenbacher, & Sturgill, 1977; Clare & Lewandowsky, 2004; Clark, 2003, 2005; Malpass & Devine, 1981b), but also for a wide range of perceptual and decision tasks (see Swets, 1996).

Consistent with this two-factor analysis, when the accuracy of identifications in TP lineups was separated from response criterion with a criterion-free dependent measure, the correlation between accurate identifications and accurate nonidentifications was reliable.

Correlations between response probabilities in TP and TA lineups were large for nonidentifications and foil identifications, but very small, and not significantly different from zero for suspect identifications. This pattern of correlations is consistent with a view that nonidentifications and foil identifications are largely the product of variation in response criterion, whereas suspect identifications are highly dependent upon the accuracy of the witness's memory of the target person.

Diagnosticity of witness responses

Suspect identifications were clearly and consistently more diagnostic than any other response. Nonidentifications were somewhat less diagnostic than suspect identifications, but also consistently high across analyses. These analyses were consistent with results reported by Wells and Olson. However, the diagnosticity patterns for foil identifications and don't know responses are not as clear-cut, and it is for these responses that the present analyses showed differences with respect to the results reported by Wells and Olson. Briefly, although foil identifications were on the whole diagnostic of innocence, as was the case in the Wells and Olson analysis, the diagnosticity varied with lineup composition and reversed when foils were selected based on their similarity to the suspect. As for don't know responses, Wells and Olson suggested that they might be diagnostic of the suspect's innocence, but the present analysis showed don't know responses to have little or no probative value.

In two separate analyses we also examined patterns of diagnosticity in simultaneous and sequential lineups. The analyses showed a complex and inconsistent pattern of results. An analysis restricted only to studies that directly compared simultaneous and sequential lineups showed greater diagnosticity for suspect identifications for sequential lineups than for simultaneous lineups. Another analysis that compared simultaneous and sequential lineups in the entire data set did not show a diagnosticity difference for suspect identifications, but showed an advantage for simultaneous lineups for the diagnosticity of foil identifications and an advantage for sequential lineups for the diagnosticity of nonidentifications. As noted before, the full data-set comparison, because of the differences in TA similarity relations, needs to be viewed with considerable caution. Neither analysis was consistent with the Wells and Olson analysis which showed that foil identifications were more diagnostic for sequential than for simultaneous lineups. Their conclusion seems clearly to be the result of raw-frequency aggregation and a single study that dominated the result.

We have reported in this paper a rather large number of analyses that yielded results that often seemed inconsistent, with complex patterns of correlations, and patterns of diagnosticity that varied as a result of lineup procedures and lineup composition. In the next section we attempt to pull these results together in a very simple theoretical framework.

Memory, decision, and lineup composition

Eyewitness identification decisions are, in large part, the product of three factors: the accuracy of the memory trace, the decision processes, and the similarity relationships among lineup members. In some cases the interplay of these components leads to results that are quite straightforward. In other cases the interplay is not straightforward.

At the simplest level, response diagnosticity is about the accuracy of memory. Were it not for the accuracy of the witness's memory of the perpetrator, none of the identification responses would show any diagnosticity at all. Responses are diagnostic only because the witness remembers what the perpetrator looked like. Thus, suspect identifications are highly diagnostic because the guilty suspect matches the witness's memory of the perpetrator more closely than do foils or the innocent suspect. In similar fashion, assuming that nonidentification responses arise when all lineup members are poor matches to memory, nonidentification responses should be given more frequently for TA than for TP lineups, and indeed the high diagnosticity of nonidentifications shows that they are.

A straightforward interplay between memory and decision processes was shown in the correlations between correct identifications and correct nonidentifications. If, as we have asserted, both are based on the accuracy of memory, they should be highly correlated. However, the correlation was close to zero. We propose that the zero correlation is due to a trade-off between memory and decision processes operating in opposition. Specifically, variation across studies in the accuracy of memory should produce a positive correlation, but variation in the decision criterion (i.e., willingness to make an identification) should produce a negative correlation. Consistent with that reasoning, a measure of identification accuracy connected less to decision criterion, the conditional probability of a correct identification given any identification ($\text{suspTP}/\text{suspTP} + \text{suspTA}$), was positively correlated with the correct nonidentification rate.

We also suggested that response probabilities should be negatively correlated in TP and TA lineups to the extent that they are based on the accuracy of memory and positively correlated to the extent that they are based on witnesses' willingness to make an identification. The pattern of TP-TA correlations should reflect the mix of these positive and negative components. The obtained pattern of correlations is precisely what one should expect given the assumptions just described. The correlations ordered from most to least positive were don't know responses (.750), foil identifications (.559), nonidentifications (.466) and suspect identifications (.173).

If suspect identifications were based only on the accuracy of memory, with no variation due to the willingness to make an identification, the correlation would be negative. With increases in the accuracy of memory correct suspect identifications would increase and false suspect identifications would decrease. The low TP-TA correlation for suspect identifications suggests a mix of variation in the accuracy of memory and variation in the willingness to make an identification. Nonidentification responses are both a measure of a witness's willingness to make an identification and a witness's ability to correctly recognize the absence of the target. Foil identifications, because they are always errors, may be viewed as an indication of the witness's desire to make an identification despite having a weak memory of the target, thus leading to the high TP-TA correlation. Don't know responses, taken literally as an indicator of the witness's insufficient memory of the perpetrator, showed the highest positive TP-TA correlation. In this view, foil identifications and don't know responses are both given by witnesses with poor

memories for the perpetrator. The difference is simply that witnesses who identify foils are more willing to make an identification.

The effect of lineup similarity is also straightforward at least for suspect identifications. In studies which used the same foils in TP and TA lineups, the only difference between the lineups was due to the innocent suspect. As the similarity of the innocent suspect to the perpetrator increases, TP and TA lineups become less distinguishable, and response diagnosticities should decrease. The comparison of designated-innocent and estimated-innocent lineups showed precisely this result, for both suspect and foil identifications.

The least straightforward mix of memory, decision, and lineup composition is shown in the variation for foil identifications across different procedures for selecting foils for TP and TA lineups. The principle that diagnosticity should decrease as TP and TA lineups become more similar suggests that the diagnosticity of foil identifications should be lowest when the same foils were used in TP and TA lineups and highest when different foils were used, as would be the case for suspect-matched lineups. Clearly this pattern was not shown in the data. The results showed just the opposite, with high diagnosticity when the same foils were used in TP and TA lineups, and very little diagnosticity for foils in suspect-matched lineups. An explanation of these results is offered below.

Foils were selected according to three scenarios which are illustrated in the top panels of Fig. 2. Each panel represents on the vertical axis the expected match to memory for suspects and foils in TP and TA lineups. The match values in TP lineups are the same across the three scenarios, and are included in the figure as a comparison point for the TA lineups.

Panel A shows the case for estimated-innocent lineups in which the foils were constant across TP and TA lineups and the innocent suspect was assumed to be equivalent to the foils. Panel B shows the case for designated-innocent lineups in which the foils were constant across TP and TA lineups, but the innocent suspect was selected to be more similar to the target than were the foils. Panel C shows the case for suspect-matched lineups.

The variation in foil diagnosticity depends on two factors shown in Fig. 2, the expected match values for foils and the match values for foils relative to the suspect. When the foils are constant in TP and TA lineups (Panels A and B), the expected match values for the foils is the same. The only difference between Panels A and B is that in Panel B the higher similarity of the innocent suspect means that the innocent suspect is a stronger competitor than when suspect and foils are assumed equivalent (Panel A). Thus, the similarity relations in Panel B predict that suspect identifications should increase and foil identifications should decrease relative to the relationships shown in Panel A. The decrease in foil identification rates for TA lineups, with foil identification rates constant for TP lineups produces the result observed from these analyses: the diagnosticity of foil identifications decreases as the similarity of the innocent suspect is increased (as was the case in the designated-innocent lineups).

Panel C shows the case in which the innocent suspect is chosen to be similar to the perpetrator, and the foils for both TP and TA lineups are selected to be similar to the suspect in their respective lineups. Note the arrows indicating that the difference in the expected matches for suspect and foils are the same for TP and TA lineups. Because the difference in expected matches is the same for TP and TA lineups, it follows that the expected match for TA foils must be lower than the expected match for TP foils. The lower match value for TA foils than for TP foils suggests that TA foils will be identified less often than TP foils. This is precisely the result that was shown for suspect-matched lineups.

The foil identification rates for each of the three conditions—same-foils/estimated innocent, same-foils/designated innocent, and suspect-matched foils—are plotted beneath the panels of the figure that illustrate the similarity relations for that condition. The results correspond exactly to the similarity relations illustrated in the figure. Thus, the combination of two factors—the match

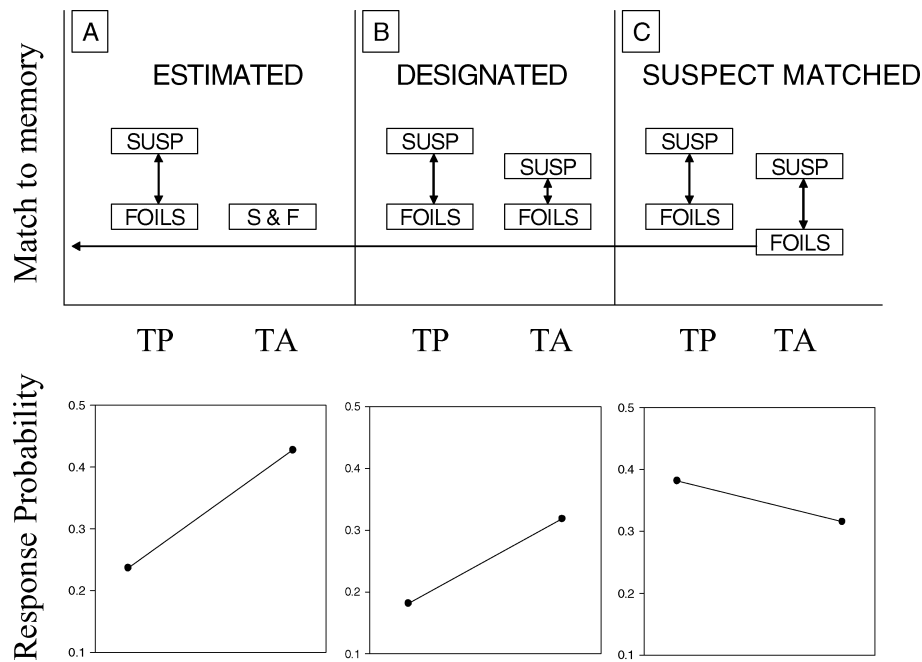


Fig. 2 Illustration of three methods of constructing target-absent lineups in eyewitness identification experiments. **A** Estimated. No distinction is made between innocent suspect and foils (denoted as S & F). **B** Designated. An innocent suspect is designated a priori, and the same foils are used for TP and TA lineup. **C** Suspect-Matched. Innocent suspect is designated a priori, and foils are selected based on match to guilty suspect in TP lineup or innocent suspect in TA lineup. Figures under each panel show the pattern of results from the meta-analysis for each lineup methodology. To equate conditions, only data from six-person, simultaneous lineups were included

values for TA foils, and the difference between TA foils and the innocent suspect—predicts the pattern of foil identifications shown in the data.

We turn now to the diagnosticity analyses for simultaneous and sequential lineups. The complex pattern of results may also be the result of a complex interplay between lineup composition and decision processes. Specifically, when simultaneous and sequential lineups were compared for the entire corpus of studies, sequential lineups showed greater diagnosticity for nonidentifications than did simultaneous lineups, and simultaneous lineups showed greater diagnosticity for foil identifications. However, when the comparison was restricted only to studies that made a direct comparison between simultaneous and sequential lineups, sequential lineups showed greater diagnosticity for suspect identifications than did simultaneous lineups. A comparison between simultaneous lineups generally and simultaneous lineups in direct comparison studies suggested that the lineups used in direct comparison studies were more biased against the innocent suspect than was the case for the set of studies that were not used in a direct simultaneous-sequential comparison. This suggests that the relative similarity between the innocent suspect and the foils in TA lineups may be an important key to the interaction.

With that assumption, the results of these two simultaneous-sequential analyses might be summarized as follows: For more biased lineups, sequential lineups showed an advantage in terms of the diagnosticity of suspect identifications, whereas for less biased lineups, the sequential advantage shifted to greater diagnosticity of nonidentifications. This appraisal of the results

should be taken with caution as it is not yet clear why the diagnosticity patterns would shift in this way.

Implications for experimental eyewitness identification research

The present analyses, which showed a near-zero correlation between correct identifications and correct nonidentifications, lead to a clear conclusion regarding TP and TA lineups: *One cannot extrapolate from TP results to TA outcomes, and vice versa.* Some factors may have a large effect on correct identification rates, but little effect on false identification rates, and vice versa. This point has been made by others, most notably, by Ebbesen and Konecni (1996).⁴ We echo their point with the results shown here.

These analyses also showed how the patterns of results of eyewitness identification experiments can be affected by two decisions that researchers must make for every experiment: how to choose an innocent suspect and how to choose the foils. A large number of studies, in fact, did not choose an innocent suspect, but rather treated all lineup members in TA lineups as equivalent. Studies that did designate an innocent suspect appear to have often chosen one with very high similarity to the perpetrator. The implication is that people who are falsely suspected often bear an unusual resemblance to the actual perpetrator. As this is a question about what “really” happens, it is best answered by an archival analysis. The use of high-similarity innocent-suspects that show high false identifications rates may also be misleading to the extent that it suggests that false identifications of the innocent occur because of their high similarity to the actual perpetrator. As Clark and Tunnicliff (2001) have noted, this is not the case. The false identification rate for an innocent suspect may be quite high, even if he does not look that much like the perpetrator, provided that he looks more like the perpetrator than do the foils.

These analyses also suggest a disconnect between laboratory and actual criminal investigations in that the large majority of experiments use the same foils in both TP and TA lineups. In contrast, suspect-matched foil selection, the standard procedure as reported by two law enforcement surveys, produce TP and TA lineups that do not have the same foils, and are inherently biased against the innocent suspect. Experimental lineups that are assumed unbiased as the result of the estimation procedure required when there is no designated innocent suspect, and which share the same foils in TP and TA lineups, may be quite unlike the lineups typically used by law enforcement.

And finally, the analyses presented here underscore the importance of a full reporting of eyewitness identification data. It is the variation in the pattern of responses that is important, rather than any particular response probability. In addition, it is important that experimental methodologies and written reports distinguish between foil identifications versus the identification of a designated innocent suspect. More importantly, our analysis strongly suggests that estimating the likelihood of a misidentification of an innocent suspect by dividing the total number of identifications by the lineup size is likely to underestimate the true rate of such misidentifications from suspect-matched lineups used by law enforcement.

⁴Ebbesen and Konecni (1996) directed their criticism at the failure to distinguish between misses versus false alarms. Their use of the term *false alarm* includes identifications of foils in both TP and TA lineups, as well as identifications of innocent suspects in TA lineups. Our concern here is directed at errors in TP lineups versus TA lineups; however, the point is the same—that experts should not fail to distinguish between different categories of identification errors, as they are likely to be produced by different mechanisms, and are certainly to have different legal implications.

The meaning of eyewitness identification decisions

Eyewitness identification decisions are interpreted by a long list of participants in the criminal justice system, from the police officer who obtains the evidence, to the defense and prosecuting attorneys, the trial judge, the jury, and if the defendant is convicted, the participants in post-conviction appeals. The question is: What does the witness's response mean?

The clearest case is for suspect identifications. An identification of the suspect is diagnostic and therefore evidence of the suspect's guilt. However, the present analyses suggest that a suspect identification is less informative if the lineup is biased. The present analyses also suggest that in a biased lineup, suspect identifications have less probative value in a simultaneous lineup than in a sequential lineup. Based on previous analyses (Clark, 2005; Steblay, 1997) we would add that the probative value of a suspect identification is undermined as well if lineup instructions increase the witness's willingness to make an identification. A general principle that emerges from these analyses is that a suspect identification has greater probative value to the extent that it is based on the witness's memory, and less probative value to the extent that it is due to lineup composition or an increase in the witness's conformity, willingness, or desire to make an identification.

Nonidentifications also are straightforward. They are diagnostic of the suspect's innocence. We reiterate the point made by Wells and Lindsay (1980) that nonidentifications are not merely "failures" to identify the suspect, but rather carry important information whose value should not be overlooked. It is important to note as well that lineup rejections carry a different meaning than don't know responses. The distinction between don't know and reject responses is important. In contrast to the witness who responds don't know, the witness who rejects the lineup may be more clearly stating "I *do* know—that the culprit is *not* in the lineup."

The least straightforward case is for foil identifications. What does a foil identification bring to the question of the suspect's guilt? Wells and Olson (2002), based on their analyses, concluded that foil identifications were indicators of the suspect's innocence. Our analysis suggests that this may be true if the lineup foils are selected based on their match to the witness's description of the perpetrator, but not if the foils are selected based on their match to the suspect. In the suspect-matched case, foil identifications appear either to have no probative value, or to be indicative of the suspect's guilt.

Admissibility and scope of expert testimony

The admissibility of expert testimony on eyewitness identification requires a reliable empirical foundation. First, we note at the outset, that contrary to the opinion in *N.Y. v. Smith*, there are dozens of studies which allow a direct TP-TA lineup comparison. In addition to the 94 comparisons here, there are many more that were excluded for reasons quite unrelated to their relevance to the justice system.

Of course, it is not sufficient for there simply to be lots of data; those data must be reliable, and must allow a coherent understanding of underlying mechanisms. In order for expert testimony to be admissible under the *Federal Rules of Evidence* (U.S. Congress, 2004) and *Daubert* criteria, the empirical foundation must be reliable, showing consistent patterns of results in order to assist rather than confuse the jury. Certainly, if one were to consider only the wide range of results, say for correct identifications (8 to 80%) or for misidentifications (0 to 70%), one might wonder if laboratory experiments can tell the court anything about the reliability of eyewitness identification. The present results show that through what might appear as very inconsistent, highly variable results, consistent response patterns and consistent, predictable variations in

those response patterns, emerge. We emphasize again that the regularities in eyewitness identification are shown in the patterns of responses and their co-variation, rather than in the response probabilities themselves.

Conclusions and future research

Our suggestions for future research arise from those sections of the paper in which, en route to selecting the “right” studies for particular comparisons, we ended up with a very small handful of studies. In particular, there is very little research using lineups in which foils are selected based on their similarity to the suspect. Given survey results showing that this is the most common procedure used by police for selecting lineup foils, it is imperative that the procedure be incorporated into laboratory research as well. Of course, many have argued that it is a flawed procedure (Clark & Tunnicliff, 2001; Navon, 1992; Wells et al., 1993, 1998). However, this is an argument as to why it requires *more* study, rather than less, as the suspect-matched lineup procedure may be an intuitively right-sounding procedure that may actually produce very high rates of false identification.

The other corner in which the number of studies became small was in comparing simultaneous and sequential lineups. The comparison between simultaneous lineups in general and those used in comparisons to sequential lineups suggests that simultaneous-sequential comparisons may use lineups that are more biased than those in the broader literature. This suggests that we do not know enough about simultaneous and sequential lineups when the lineup composition is less biased. Progress on these fronts is likely to require the development and application of more precise theories of the underlying memory and decision processes. Recently, some researchers have begun to develop computational models of the changes in eyewitness memory over time (Gronlund, 2005) as well as the specifics of the decision processes that underlie eyewitness identification decisions (Clare & Lewandowsky, 2004; Clark, 2003). Future theory testing and development should be guided and constrained by the patterns of regularity and variability shown for eyewitness identification in the present paper.

Appendix

Results from all 94 studies, suspect, foil, and nonidentification responses, for target-present and target-absent lineups are shown in Table A.1. An explanation of the various abbreviations is given in Table A.2.

Table A.1 Identification response probabilities for 94 studies used in the meta-analysis

Study	Target present			Target absent			
	Sim/Seq	Susp	Foil	NoID	Susp	Foil	NoID
Behrman and Richards (2005)	sim	0.676	0.054	0.270	0.055	0.276	0.669
Brewer et al. (2002)	sim	0.363	0.150	0.487	0.101	0.173	0.727
Clare and Lewandowsky (C) (2004)	sim	0.803	0.133	0.067	0.045	0.728	0.227
Clare and Lewandowsky (F) (2004)	sim	0.692	0.115	0.192	0.000	0.480	0.520
Clare and Lewandowsky (H) (2004)	sim	0.574	0.064	0.362	0.200	0.280	0.520
Clark and Tunnicliff (2001)	sim	0.339	0.468	0.194	0.254	0.159	0.587
Cutler et al. (1987a, 1987b)	seq	0.431	0.514	0.056	0.085	0.593	0.323
Dekle et al. (1996)	sim	0.300	0.140	0.560	0.060	0.240	0.700
Devenport and Fisher (auth/B) (1996)	sim	0.550	0.310	0.130	0.133	0.667	0.200
Devenport and Fisher (auth/U) (1996)	sim	0.240	0.290	0.470	0.078	0.392	0.530

Table A.1 Continued

Study	Target present			Target absent			
	Sim/Seq	Susp	Foil	NoID	Susp	Foil	NoID
Devenport and Fisher (no auth/B) (1996)	sim	0.280	0.280	0.440	0.118	0.592	0.290
Devenport and Fisher (no auth/U) (1996)	sim	0.290	0.180	0.530	0.078	0.392	0.530
Dunning and Stern (Ex 4) (1994)	sim	0.360	0.120	0.520	0.179	0.375	0.446
Dysart and Lindsay (NQ) (2001)	sim	0.650	0.000	0.350	0.083	0.417	0.500
Dysart and Lindsay (Q) (2001)	sim	0.600	0.000	0.400	0.033	0.167	0.800
Fleet et al. (neg) (1987)	sim	0.542	0.208	0.250	0.076	0.379	0.546
Fleet et al. (neu) (1987)	sim	0.652	0.261	0.087	0.097	0.486	0.417
Fleet et al. (pos) (1987)	sim	0.625	0.167	0.208	0.093	0.467	0.440
Geiselman et al. (delay) (1993)	sim	0.095	0.476	0.429	0.058	0.292	0.650
Geiselman et al. (immed) (1993)	sim	0.119	0.476	0.405	0.066	0.329	0.605
Gonzalez et al. (Ex 1) (1993)	sim	0.429	0.214	0.357	0.125	0.375	0.500
Gonzalez et al. (Ex 2) (1993)	sim	0.133	0.117	0.750	0.083	0.117	0.800
Juslin et al. (dm) (1996)	sim	0.520	0.110	0.380	0.090	0.120	0.780
Juslin et al. (sm) (1996)	sim	0.440	0.200	0.350	0.090	0.170	0.730
Krafka and Penrod (2 hr C) (1985)	sim	0.600	0.000	0.400	0.033	0.167	0.800
Krafka and Penrod (2 hr NC) (1985)	sim	0.273	0.091	0.636	0.017	0.083	0.900
Krafka and Penrod (24 hr C) (1985)	sim	0.500	0.300	0.200	0.083	0.417	0.500
Krafka and Penrod (24 hr, NC) (1985)	sim	0.308	0.154	0.538	0.091	0.454	0.455
Kassin et al. (1991)	sim	0.475	0.300	0.225	0.111	0.555	0.333
Kneller et al. (2001)	seq	0.500	0.111	0.389	0.111	0.111	0.778
Kneller et al. (2001)	sim	0.611	0.167	0.222	0.278	0.333	0.389
Lindsay (1986)	sim	0.670	0.190	0.150	0.190	0.420	0.380
Lindsay et al. (1991)	seq	0.467	0.067	0.400	0.054	0.117	0.767
Lindsay et al. (1991)	sim	0.567	0.200	0.233	0.200	0.367	0.433
Lindsay et al. (1991)	seq	0.767	0.033	0.200	0.033	0.033	0.934
Lindsay et al. (1991)	sim	0.667	0.039	0.300	0.200	0.100	0.700
Lindsay et al. (biased) (1987)	sim	0.700	0.140	0.160	0.380	0.170	0.450
Lindsay et al. (same) (1987)	sim	0.690	0.100	0.200	0.100	0.160	0.740
Lindsay et al. (unbiased) (1987)	sim	0.650	0.100	0.250	0.210	0.210	0.580
Lindsay and Wells (1985)	seq	0.500	0.020	0.480	0.170	0.180	0.650
Lindsay and Wells (1985)	sim	0.580	0.120	0.300	0.430	0.150	0.420
Lindsay and Wells (BLC) (1980)	sim	0.710	0.120	0.180	0.700	0.040	0.260
Lindsay and Wells (ULC) (1980)	sim	0.580	0.290	0.130	0.310	0.410	0.280
Maclin et al. (2005)	sim	0.433	0.333	0.233	0.094	0.472	0.433
Maclin et al. (2005)	seq	0.300	0.167	0.533	0.053	0.264	0.683
Malpass and Devine (B) (1981a, 1981b)	sim	0.750	0.250	0.000	0.156	0.624	0.220
Malpass and Devine (U) (1981a, 1981b)	sim	0.830	0.000	0.170	0.067	0.266	0.667
Melara et al. (Ex 1) (1989)	seq	0.125	0.250	0.625	0.042	0.208	0.750
Melara et al. (Ex 1) (1989)	sim	0.250	0.625	0.125	0.167	0.833	0.000
Memon et al. (LE) (2003)	sim	0.900	0.075	0.025	0.076	0.379	0.545
Memon et al. (SE) (2003)	sim	0.320	0.435	0.245	0.142	0.708	0.150
Murray and Wells (inf) (1982)	sim	0.430	0.130	0.430	0.110	0.350	0.540
Murray and Wells (uninf) (1982)	sim	0.260	0.370	0.370	0.110	0.260	0.630
Paley and Geiselman (B) (1989)	sim	0.530	0.400	0.070	0.150	0.750	0.100
Paley and Geiselman (U) (1989)	sim	0.400	0.200	0.400	0.067	0.333	0.600
Paley and Geiselman (Ex 1) (1989)	sim	0.433	0.200	0.367	0.111	0.556	0.333
Parker and Carranza (1989)	sim	0.080	0.250	0.670	0.170	0.170	0.670
Parker and Carranza (1989)	sim	0.420	0.250	0.330	0.330	0.420	0.250
Parker and Ryan (1993)	seq	0.083	0.167	0.750	0.083	0.167	0.750

Table A.1 Continued

Study	Target present			Target absent			
	Sim/Seq	Susp	Foil	NoID	Susp	Foil	NoID
Parker and Ryan (1993)	sim	0.417	0.167	0.417	0.250	0.333	0.417
Pozzulo et al. (1999)	sim	0.800	0.000	0.200	0.000	0.130	0.870
Read (HS, LE) (1995)	sim	0.467	0.467	0.068	0.083	0.334	0.583
Read (HS, SE) (1995)	sim	0.385	0.308	0.308	0.027	0.106	0.867
Read (LS, LE) (1995)	sim	0.273	0.364	0.364	0.093	0.374	0.533
Read (LS, SE) (1995)	sim	0.176	0.117	0.706	0.100	0.400	0.500
Read et al. (Ex 3) (1990)	sim	0.800	0.100	0.100	0.060	0.400	0.540
Sanders and Simmons (WH) (1983)	sim	0.136	0.273	0.591	0.300	0.400	0.300
Sanders and Simmons (WNH) (1983)	sim	0.429	0.238	0.333	0.125	0.375	0.500
Searcy et al. (1999)	sim	0.263	0.421	0.316	0.026	0.605	0.368
Semmler et al. (2004)	sim	0.337	0.159	0.504	0.028	0.193	0.779
Smith et al. (cross race) (2001)	sim	0.453	0.172	0.375	0.154	0.354	0.492
Smith et al. (same race) (2001)	sim	0.458	0.136	0.407	0.066	0.246	0.689
Smith et al. (cross-race) (2004)	sim	0.368	0.211	0.421	0.064	0.511	0.426
Smith et al. (same race) (2004)	sim	0.475	0.125	0.400	0.056	0.333	0.611
Sporer (1992)	sim	0.500	0.094	0.406	0.100	0.367	0.533
Sporer (1993)	seq	0.389	0.167	0.444	0.102	0.509	0.389
Sporer (1993)	sim	0.444	0.333	0.222	0.120	0.602	0.278
Tunnicliff and Clark (Ex 1 dm) (2000)	sim	0.531	0.156	0.313	0.125	0.343	0.531
Tunnicliff and Clark (Ex 1 sm) (2000)	sim	0.531	0.250	0.218	0.031	0.313	0.655
Tunnicliff and Clark (Ex 2 dm) (2000)	sim	0.313	0.250	0.438	0.188	0.354	0.458
Tunnicliff and Clark (Ex 2 sm) (2000)	sim	0.333	0.271	0.396	0.083	0.188	0.729
Wells (1984)	sim	0.604	0.208	0.188	0.312	0.396	0.292
Wells et al. (1981)	sim	0.570	0.250	0.180	0.480	0.330	0.200
Wells et al. (Comp) (2005)	sim	0.180	0.200	0.620	0.043	0.217	0.740
Wells et al. (No Comp) (2005)	sim	0.600	0.040	0.360	0.033	0.167	0.800
Wells et al. (dm) (1993)	sim	0.667	0.071	0.262	0.119	0.310	0.572
Wells et al. (mm) (1993)	sim	0.700	0.080	0.230	0.480	0.150	0.360
Wells et al. (sm) (1993)	sim	0.214	0.429	0.357	0.119	0.476	0.405
Wright and Stroud (2002)	sim	0.400	0.380	0.220	0.090	0.558	0.350
Yarmey et al. (1994)	sim	0.460	0.240	0.290	0.050	0.230	0.720
Yarmey et al. (2 hr) (1996)	hyb	0.360	0.460	0.180	0.140	0.550	0.310
Yarmey et al. (24 hr) (1996)	hyb	0.320	0.470	0.210	0.140	0.570	0.290
Yarmey et al. (30 min) (1996)	hyb	0.390	0.350	0.260	0.330	0.390	0.280
Yarmey et al. (immed) (1996)	hyb	0.490	0.280	0.230	0.160	0.460	0.380

Table A.2 Key to abbreviations in Table A.1

Study or studies	Abbreviation key
Clare and Lewandowsky (2004)	F, H, C = Featural, holistic encoding, or control condition
Devenport and Fisher (1996)	Auth, No Auth = lineup administrator dressed as authority figure, or not dressed as authority figure; B, U = biased and unbiased lineup instructions
Dysart and Lindsay (2001)	Q = prelineup question, NQ = no prelineup question.
Fleet et al. (1987)	neu, pos, neg = neutral, positive biased and negative biased instructions
Juslin et al. (1996)	sm, dm, mm = suspect-matched,

Table A.2 Continued.

Study or studies	Abbreviation key
Tunnicliff and Clark (2000)	description-matched, and mismatched foil selection
Wells et al. (1993)	
Krafka and Penrod (1985)	2 hr, 24 hr delays, C = context reinstated, NC = context not reinstated
Lindsay et al. (1987)	BC, UC, SC = biased clothing, unbiased clothing, and same clothing
Malpass and Devine (1981b)	B, U - biased or unbiased lineup instructions
Paley and Geiselman (1989)	
Memon et al. (2003)	SE, LE = short exposure, long exposure
Murray and Wells (1982)	inf = witnesses informed that crime was staged uninf = witnesses uninformed that crime was staged
Read (1995)	HS, LS = high- or low-similarity lineup; SE, LE = short exposure, long exposure
Sanders and Simmons (1983)	WH = witnesses hypnotized WNH = witnesses not hypnotized
Wells and Lindsay (1980)	blc = biased lineup composition, ulc = unbiased lineup composition
Wells et al. (2005)	Comp, No Comp = witnesses either presented or not presented with composite drawing before lineup.
Yarmey et al. (1996)	immed, 2 hr, 24 hr, or 30 min delays between exposure and identification

Acknowledgements This research was supported by National Science Foundation grant SES 0214373 awarded to Steven Clark. We wish to thank Brynn Nodarse for her assistance, and to all the reviewers for their very helpful comments and suggestions which improved the paper substantially.

References

- *Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied*, 8, 44–56.
- Brown, E., Deffenbacher, K., & Sturgill, W. (1977). Memory for faces and the circumstances of encounter. *Journal of Applied Psychology*, 62, 311–318.
- *Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 4, 739–755.
- Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology*, 17, 629–654.
- Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law & Human Behavior*, 29, 395–424.
- Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models fit the data. *Psychonomic Bulletin and Review*, 3, 37–60.
- *Clark S. E., & Tunnicliff, J. L. (2001). Selecting lineup foils in eyewitness identification experiments: Experimental control and real-world simulation. *Law & Human Behavior*, 25, 199–216.
- *Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987a). Improving the reliability of eyewitness identification: Putting context into context. *Journal of Applied Psychology*, 72, 629–637.
- *Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987b). The reliability of eyewitness identification: The role of system and estimator variables. *Law & Human Behavior*, 11, 233–258.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

*Included in the meta-analysis.

- Daubert et al., v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
- Deffenbacher, K. A., Bornstein, B. H., & Penrod, S. D. (2006). Mugshot exposure effects: Retroactive interference, mugshot commitment, source confusion, and unconscious transference. *Law & Human Behavior*, 30, 287–307.
- Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law & Human Behavior*, 6, 687–706.
- *Dekle, D. J., Beal, C. R., Elliott, R., & Huneycutt, D. (1996). Children as witnesses: A comparison of lineup versus showup identification methods. *Applied Cognitive Psychology*, 10, 1–12.
- *Devenport J. L., & Fisher R. P. (1996). The effects of authority and social influence on eyewitness suggestibility and person recognition. *Journal of Police and Criminal Psychology*, 11, 35–40.
- *Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, 67, 818–835.
- *Dysart, J. E., & Lindsay, R. C. L. (2001). A preidentification questioning effect: Serendipitously increasing correct rejections. *Law & Human Behavior*, 25, 155–165.
- Ebbesen, E. B., & Konecni, V. J. (1996). Eyewitness memory research: Probative v. prejudicial value. *Expert Evidence: The international Digest of Human Behaviour, Science, and the Law*, 5, 2–28.
- Egeth, H. E. (1993). What do we *not* know about eyewitness identification. *American Psychologist*, 48, 577–580.
- Egeth, H. E. (1995). Expert psychological testimony about eyewitnesses: An update. In F. Kessel (Ed.) *Psychology, science, and human affairs: Essays in honor of William Bevan* (pp. 151–166). Boulder, CO: Westview Press.
- Elliott, R. (1993). Expert testimony about eyewitness identification: A critique. *Law & Human Behavior*, 4, 423–437.
- *Fleet, M. L., Brigham, J. C., & Bothwell, R. K. (1987). The confidence-accuracy relationship: The effects of confidence assessment and choosing. *Journal of Applied Social Psychology*, 17, 171–187.
- *Geiselman, R. E., MacArthur, A., & Meerovitch, S. (1993). Transference of perpetrator roles in eyewitness identifications from photoarrays. *American Journal of Forensic Psychology*, 11, 5–15.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: Wiley.
- Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, 11, 8–20.
- Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 1, 5–16
- *Gonzalez, R., Ellsworth, P. C., & Pembroke, M. (1993). Response biases in lineups and show ups. *Journal of Personality & Social Psychology*, 4, 525–537.
- Gronlund, S. D. (2005). Sequential lineup advantage: Contributions of distinctiveness and recollection. *Applied Cognitive Psychology*, 19, 23–37.
- Hintzman, D. L. (1980). Simpsons Paradox and the analysis of memory retrieval. *Psychological Review*, 87, 398–410.
- Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers*. Milwaukee, WI: ASQC Quality Press.
- *Justin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 22, 1304–1316.
- *Kassin, S. M., Rigby, S., & Castillo, S. R. (1991). The accuracy-confidence correlation in eyewitness testimony: Limits and extensions of the retrospective self-awareness effect. *Journal of Personality & Social Psychology*, 61, 689–707.
- *Kneller, W., Memon, A., & Stevenage, S. (2001). Simultaneous and sequential lineups: Decision processes of accurate and inaccurate eyewitnesses. *Applied Cognitive Psychology*, 15, 659–671.
- *Krafka, C., & Penrod, S. (1985). Reinstatement of context in a field experiment on eyewitness identification. *Journal of Personality & Social Psychology*, 49, 58–69.
- *Lindsay, R. C. (1986). Confidence and accuracy of eyewitness identification from lineups. *Law & Human Behavior*, 10, 229–239.
- *Lindsay, R. C. L., Lea, J. A., & Fulford, J. A. (1991). Sequential lineup presentation: Technique matters. *Journal of Applied Psychology*, 76, 741–745.
- *Lindsay R. C. L., Lea J. A., Nosworthy G. J., Fulford J. A., Hector J., LeVan V., & Seabrook C. (1991). Biased lineups: Sequential presentation reduces the problem. *Journal of Applied Psychology*, 76, 796–802.
- Lindsay, R. C. L., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions: a problem for the match-to-description lineup foil selection strategy. *Law & Human Behavior*, 18, 527–541.
- *Lindsay, R. C., Wallbridge, H., & Drennan, D. (1987). Do the clothes make the man? An exploration of the effect of lineup attire on eyewitness identification accuracy. *Canadian Journal of Behavioural Science. Special Forensic Psychology*, 19, 463–478.
- *Lindsay, R. C., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law & Human Behavior*, 4, 303–313.

- *Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*, 556–564.
- Malpass, R. S., & Devine, P. G. (1981a). Guided memory in eyewitness identification. *Journal of Applied Psychology, 3*, 343–350.
- *Malpass, R. S., & Devine, P. G. (1981b). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology, 4*, 482–489.
- McCloskey, M., & Egeth, H. (1983). Eyewitness identification: What can a psychologist tell a jury? *American Psychologist, 38*, 550–563.
- Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition, 33*, 783–792.
- *Melara, R. D., DeWitt-Rickards, T. S., & O'Brien, T. P. (1989). Enhancing lineup identification accuracy: Two codes are better than one. *Journal of Applied Psychology, 74*, 706–713.
- Memon, A., Hope, L., & Bull, R. (2003). Exposure duration: Effects on eyewitness accuracy and confidence. *British Journal of Psychology, 3*, 339–354.
- *Murray, D. M., & Wells, G. L. (1982). Does knowledge that a crime was staged affect eyewitness performance? *Journal of Applied Social Psychology, 12*, 42–53.
- Navon, D. (1992). Selection of lineup foils by similarity to suspect is likely to misfire. *Law and Human Behavior, 16*, 575–593.
- *Paley, B., & Geiselman, R. E. (1989). The effects of alternative photospread instructions on suspect identification performance. *American Journal of Forensic Psychology, 7*, 3–13.
- *Parker, J. F., & Carranza, L. E. (1989). Eyewitness testimony of children in target-present and target-absent lineups. *Law & Human Behavior, 13*, 133–149.
- *Parker, J. F., & Ryan, V. (1993). An attempt to reduce guessing behavior in children's and adults' eyewitness identifications. *Law & Human Behavior. Special Law, Psychology, and Children, 17*, 11–26.
- People v. Smith* (2004). 2 Misc.3d 1007(A), N.Y. Slip Op. 50172(U).
- *Pozzulo, J. D., & Lindsay, R. C. L. (1999). Elimination lineups: An improved identification procedure for child eyewitnesses. *Law & Human Behavior, 84*, 167–176.
- Py, J., Demarchi, S., Ginot, M., & Wasiak, L. (2003). Who is the suspect? A complementary instruction to the standard mock witness paradigm. Presented at Psychology and Law: International, Interdisciplinary Conference, Edinburgh, Scotland.
- Read, J. D. (1995). The availability heuristic in person identification: The sometimes misleading consequences of enhanced contextual information. *Applied Cognitive Psychology, 2*, 91–121.
- *Read J. P., Tollestrup P., Hammersley R., McFadzen E., & Christensen A. (1990). The unconscious transference effect: Are innocent bystanders ever misidentified? *Applied Cognitive Psychology, 4*, 3–31.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park: Sage Publications.
- *Sanders, G. S., & Simmons, W. L. (1983). Use of hypnosis to enhance eyewitness accuracy: Does it work? *Journal of Applied Psychology, 68*, 70–77.
- *Searcy, J. H., Bartlett, J. C., & Memon, A. (1999). Age differences in accuracy and choosing in eyewitness identification and face recognition. *Memory & Cognition, 27*, 538–552.
- Semmler, C., Brewer, N., & Wells, G. L. (2004). Effects of postidentification feedback on eyewitness identification and nonidentification confidence. *Journal of Applied Psychology, 89*, 334–346.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, B, 13*, 238–241.
- *Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology, 78*, 22–33.
- *Sporer, S. L. (1992). Post-dicting eyewitness accuracy: Confidence, decision-times and person descriptions of choosers and non-choosers. *European Journal of Social Psychology, 74*, 157–180.
- Stebly, N. M. (1992). A meta-analytic review of the weapon focus effect. *Law & Human Behavior, 4*, 413–424.
- Stebly, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law & Human Behavior, 5*, 459–473.
- Stebly, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law & Human Behavior, 5*, 523–540.
- Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Starr, S. A., & Williams, R. M., Jr. (1949). *The American soldier: Adjustment during army life* (vol. I). Princeton, NJ: Princeton University Press.
- Swets, J. A. (1996). *Signal detection theory and RO analysis in psychology and diagnosis*. Mahwah, NJ: Erlbaum.
- *Tunnicliff, J. L., & Clark, S. E. (2000). Selecting foils for identification lineups: Matching suspects or descriptions? *Law & Human Behavior, 24*, 231–258.
- U.S. Congress. (2004). *Federal Rules of Evidence*. Committee on the Judiciary, 108th Congress, House of Representatives. Washington, DC: U.S. Government Printing Office.

- *Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14*, 89–103.
- Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist, 48*, 553–571.
- *Wells, G. L., Charman, S. D., & Olson, E. A. (2005). Building face composites can harm lineup identification performance. *Journal of Experimental Psychology: Applied, 11*, 147–156.
- *Wells, G. L., Ferguson, T. J., & Lindsay, R. C. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology, 66*, 688–696.
- Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 3*, 776–784.
- Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied, 3*, 155–167.
- *Wells, G. L., Rydell, S. M., & Seclau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*, 835–844.
- Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law & Human Behavior, 22*, 603–647.
- Wells, G. L., & Turtle, J. W. (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin, 99*, 320–329.
- Wogalter, M. S., Marwitz, D. B., & Leonard, D. C. (1992). Suggestiveness in photo spread line-ups: Similarity induces distinctiveness. *Applied Cognitive Psychology, 5*, 443–453.
- Wogalter, M. S., Malpass, R. S., & Burger, M. A. (1993). How police officers construct lineups: A national survey. *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, pp. 640–644.
- Wogalter, M. S., Malpass, R. S., & McQuiston, D. E. (2004). A national survey of U.S. police on preparation and conduct of identification lineups. *Psychology, Public Policy, & Law, 10*, 69–82.
- *Wright, D. B., & Stroud, J. N. (2002). Age differences in lineup identification accuracy: People are better with their own age. *Law & Human Behavior, 66*, 641–654.
- *Yarmey, A. D., Yarmey, M. J., & Yarmey, A. L. (1996). Accuracy of eyewitness identification in showups and lineups. *Law & Human Behavior, 4*, 459–477.
- *Yarmey, A. D., Yarmey, A. L., & Yarmey, M. J. (1994). Face and voice identifications in showups and lineups. *Applied Cognitive Psychology, 8*, 453–464.
- Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika, 2*, 121–134.