CHAPTER 13

Eyewitness Identification Research: Strengths and Weaknesses of Alternative Methods

GARY L. WELLS AND STEVEN D. PENROD

EYEWITNESS IDENTIFICATION RESEARCH has become an increasingly popular specialty among those interested in psychology and law. We have tried to write this chapter in a way that would be useful for new or "would be" eyewitness researchers, and we also think that there might be some points here that seasoned eyewitness researchers might find useful. The reader should keep in mind, however, that that there can be different views about what is the best methodology for eyewitness research. We offer our best advice, but that is no guarantee that following our advice will automatically disarm critics of any researcher's study. Furthermore, researchers should always consider any particular recommendation that we make in the context of the hypotheses, purposes, and goals of their own study.

For example, one of the authors of this chapter favors a methodology in which fillers are counterbalanced in and out of the position of the innocent suspect rather than having one person always serve as the innocent replacement for the culprit (a general preference of the second author). Both authors agree, however, that if the purpose of the study is to test ideas about lineup bias emanating from poor fillers, the counterbalancing methodology would not make sense. We cannot anticipate every hypothesis that researchers would want to pursue, so researchers will always have to think through every methodological decision within the context of the questions they are trying to answer.

Readers should also note that considerations of ecological validity, although often dominant in eyewitness research, can sometimes take a back seat to theoretical considerations. Consider, for example, the removal-without-replacement procedure for lineups in which the culprit-absent lineup is created by removing the culprit and replacing him/her with no one (Wells, 1993).
From the perspective of ecological validity, this is a peculiar methodology without much of a real-world counterpart. On the other hand, this methodology is powerful for testing certain types of theoretical questions and is, in some ways, a "cleaner" manipulation of the culprit-present lineup versus culprit-absent lineup variable. This is because the traditional method (replace the culprit with someone else) changes two variables at once (the culprit is removed, and a new face is introduced), whereas the removal-without-replacement procedure removes only the culprit. Depending on the purposes of the study, ecological validity does not always trump other concerns. In fact, the relative paucity of theory development in the eyewitness area is something that many researchers now believe needs to be rectified, which could lead to the use of methodologies that are less ecologically valid but more revealing of psychological processes (see Applied Cognitive Psychology, 2008, Volume 22(6), devoted to such discussions).

Readers should also keep in mind that methodological soundness is only one factor that determines whether an eyewitness study is likely to survive the rigorous review process and be accepted for publication. Many manuscripts are rejected from top journals despite their methodological soundness because the study does not significantly advance our knowledge. One of the poorest reasons to conduct a study is merely that "no one has done it." All studies have to be justified at much deeper levels than that. Why is it important? How does it relate to the current literature? How does it advance our theoretical understanding? What are the applied implications? Novice eyewitness researchers would do well to examine carefully the introduction and discussion sections of well-cited eyewitness articles to develop a feel for how these articles were "packaged."

**EYEWITNESS IDENTIFICATION LABORATORY
EXPERIMENTAL METHODS**

Designing, conducting, and analyzing an eyewitness identification experiment involves a series of decisions that can have considerable consequences for the results and their interpretations. Although we cannot exhaust all the possible decision junctures that confront eyewitness identification researchers, many of which depend on specifics of the hypothesis being tested, there are several that confront researchers in the design of almost all eyewitness identification experiments.

The basic idea of an eyewitness identification experiment is to expose participant witnesses to an event in which they view one or more actors engaged in some critical behavior. Later, the participant witnesses are given some type of eyewitness identification test, usually a photo array, in which they are asked to attempt to identify the individual(s). In most eyewitness identification experiments, the participant witnesses are also asked to indicate the confidence or certainty they have in their identification decisions. Depending on the focus of the experiment, other measures might be taken, such as determining if their identification decisions are made under conditions that are more realistic, such as when they view video recordings of the event as it actually occurred.
such as decision latency, willingness of participant witnesses to testify about their identification decision, statements about how they made their decisions, and so on. In many respects, such an experiment seems rather straightforward. But the apparent simplicity of preparing such an experiment is somewhat deceptive.

**The Event**

One of the initial decisions to be made is what kind of event to create. There are many examples in the eyewitness identification literature of live staged events, such as thefts, in which unsuspecting people become eyewitnesses (see Wells, Rydell, & Seelau, 1993, for an example). Increasingly, however, eyewitness identification researchers have come to rely on video events that the researchers create, often depicting some type of crime. Using video events is certainly easier than using live events, because video does not require continuing reenactments with actors, there are no concerns about participant witnesses intervening to stop the crime, the event is perfectly consistent in its presentation from one participant to the next, and institutional review boards typically have fewer concerns about video events than about live events. On the other hand, it is possible to raise concerns about ecological validity with video events due to the obvious fact that eyewitnesses in real cases witness events live, in three dimensions, and so on. But, at this point, it seems that eyewitness scientists have largely accepted the video-event method, leaving little incentive for researchers to use the more difficult and costly live-event method.

Regardless of the medium chosen for the witnessed event, we caution researchers to avoid designs in which large groups are nested within conditions of the study. At the extreme, imagine that a researcher was testing the weapon-focus effect by having a classroom disruptor flash a knife in Class A and the same disruptor have no weapon when disrupting Class B. In this case, groups (Classes A and B) are nested within conditions. Any differences between the classes are therefore confounded with the manipulated variable (weapon vs. no weapon). Surprisingly enough, the problem of nesting is not solved even if there were 400 students and they had been randomly assigned to Class A or Class B. The reason that this does not solve the problem is that it takes only one small glitch (which could easily be undetected) to affect an entire class (or a large portion of the class); if something affects the class, it affects the entire condition.

For example, maybe a student was suspicious, and a whispering rumor spread quickly in only one class. Contrast this with a fully randomized design in which the event was staged 400 times, once for each witness, and each was randomly assigned to the weapon versus no weapon condition. That one suspicious student could affect only one data point, which would have no appreciable effect on the condition in which that student fell. Consider also the possibility that the actor who plays the role of the disruptor is unable...
to enact the event in exactly the same way each time (which is likely). Sometimes the actor faces away from the group more than other times, or sometimes the actor moves a little faster. Any minor difference for Class A versus Class B could impact results for the entire condition (because each class is a full condition).

At the other extreme, where each enactment is done for each individual witness, these minor variations "wash out," because they will occur as often in one condition as they do in the other condition (the law of large numbers). Based on strict interpretations of a random design and the assumptions behind statistical analyses, the situation in which groups (Class A and Class B) are nested in conditions (weapon and no weapon) is the same as an experiment with a total sample size of two (one per condition), not 400 (200 per condition). These are extremes, of course, where in the nested case there is only one group per condition and in the other case there are no nestings of groups within conditions. Most reviewers and editors will permit some level of nesting within conditions. For example, groups of two or three witnesses at a time are usually not flagged as problematic nestings if the overall sample size is large enough. Still, researchers need to be prepared for the possibility that editors or reviewers will require some complicated higher-order statistical analyses that are harmful to statistical power when there are nestings. And, at the extreme where there is only one group per condition, there is no statistical solution. Such designs will be (and should be) rejected for publication in reputable journals.

Once a medium has been chosen for the event, questions arise, such as what participant witnesses are told prior to the event, how long the participant witnesses will be exposed to the culprit(s), and whether to use multiple culprits reenacting the same event. Most eyewitness identification researchers use mild deception to avoid cueing the witnesses prior to the event that they will witness. For example, in a video-event study, participant witnesses might be told that the study involves impressions of people or impressions of events. In live-event experiments, participants might be seated in a waiting room believing that the study has not yet started. The rationale for not alerting participant witnesses that they will be witnesses is simply that actual witnesses are generally taken by surprise.

The issue of exposure duration is a more difficult one, because the relation between exposure duration and eyewitness identification performance is not precise. The general idea is to find an exposure duration that avoids ceiling and basement effects. An extremely long exposure might result in eyewitness identification performance being virtually perfect, which might then prevent the researcher from showing that certain other variables (e.g., a better lineup procedure) can improve eyewitness identification performance. Conversely, an exposure duration that is too brief might result in such a weak memory trace that there would be no chance to show how other variables (e.g., passage of time before the lineup) can harm eyewitness identification performance.
The appropriate exposure time can vary dramatically as a function of other aspects of the research design. Thus, for example, a study of identification accuracy by military personnel being trained in a military survival school found that under some high-stress circumstances, trainees could identify just over one in four interrogators involved in 40-minute interrogations (Morgan et al., 2004). In contrast, Memon, Hope, and Bull (2003) obtained slightly better performances from witnesses who viewed a target person in a video for just 12 seconds. Hence, pilot testing is highly recommended, which can often be done using just the control condition to make sure that there is room for performance to go up and down as a function of other manipulations. (Later, we discuss how lineup performance is assessed.)

The question of whether to use multiple culprits, each reenacting the same crime, is generally construed as an issue of "stimulus sampling." The idea of stimulus sampling is to ensure that the results are not due to the particular characteristics of the individual who was selected to serve as the culprit in the study. For example, we know that some people are more recognizable than others, in large part because some people have a more distinctive (unusual) appearance than other people. However, the need for stimulus sampling is more important for some studies than it is for others. In general, the issue of stimulus sampling is extremely important when the manipulated (or predictor) variable is an exemplar from a category.

For example, if the experiment is examining how eyewitnesses perform when the culprit is Black versus White or male versus female, stimulus sampling is essential. Clearly, one must compare many different Black culprit-actors with many different White culprit-actors or many male culprit-actors with many female culprit-actors to reach any proper conclusions. Using only one or two such actori threatens construct validity (a very serious threat) in this case, because the category variable (e.g., male versus female) is confounded with characteristics of the exemplars.

However, if the critical manipulated variable is not a characteristic of the stimuli themselves (e.g., a pre-lineup instruction), then the use of only one or two culprit-actors is a less serious threat, because the characteristics of the culprit-actor are the same in each condition. In the latter case, there can be issues of generalization (i.e., would other exemplars yield the same effects from the manipulated variables?) but not issues of construct validity. We encourage readers to examine an article by Wells and Windschitl (1999) on these points; it also includes discussions of how statistical interactions can lessen the need for stimulus sampling and how to analyze the data when stimulus sampling is used.

**Building the Lineup**

Great care must be taken in making various decisions about building a lineup. The need for some decisions is obvious, such as how many people to use in the lineup, whether to present the lineup live or in photos, and so on. But
some decisions are even more complex and require considerable forethought at a level often not exercised in the eyewitness identification literature. One of the first decisions to be made is whether to use both culprit-present and culprit-absent lineups. There are circumstances in which using only present—or only absent lineups—is perfectly legitimate, depending on the nature of the question being asked. Much of the research on the post-identification feedback effect, for example, uses only culprit-absent lineups because the focus is on the development of false certainty after a mistaken identification has been made (e.g., Wells & Bradfield, 1998). Using a different rationale (a question that focused on witness performance in culprit-present arrays), Memon and Gabbert (2003) used only culprit-present lineups to examine simultaneous versus sequential lineups. Generally, however, if the hypothesis concerns eyewitness identification accuracy, it is critical to use both culprit-present and culprit-absent lineups.

The typical method for creating a culprit-absent lineup is to use the same fillers but replace the culprit with an innocent suspect. Theoretically, this simulates the situation in which police investigators have unknowingly focused the lineup on an innocent suspect. But the issue too often glossed over (or ignored altogether) by eyewitness identification researchers is how to select the person who will replace the culprit (i.e., how to select the innocent suspect). In some published experiments, researchers state that they selected a person who closely resembled the culprit. For example, the researchers might sort through a large number of photos, have them rated for similarity to the culprit, and then pick the most similar one to be the replacement for the culprit.

We believe, however, that the choice of strategy should depend on the questions being addressed by the researcher and what one assumes (if anything) about the ways in which innocent people become suspects. Most often, people do not become innocent suspects in actual cases because of their high similarity to the actual culprit. After all, the police investigators do not know who the actual culprit is (otherwise they would arrest the actual culprit instead). Instead, an innocent suspect may be someone who merely fits the general verbal description of the culprit that was given by the eyewitness. But if the innocent suspect in an experiment is selected a priori to look more like the culprit than do any of the fillers, this can serve to greatly overestimate the rate at which the innocent suspects would be identified in actual cases where the innocent person merely fits a general description.

Furthermore, selecting as the innocent replacement someone who looks a great deal like the culprit could serve to underestimate the confidence-accuracy relationship. At the extreme, replacing the culprit with a clone of the culprit would yield a result in which it is guaranteed that there can be no confidence-accuracy relation. In the general case of an experiment, we recommend a different method than one that selects a person who looks most like the culprit as the innocent replacement for the culprit. In particular, we recommend that the innocent replacement for the culprit be selected the same way that the verbal descriptor specifies any single culprit. Instead, each of the culprit-absent lineup data and make the on peculiarities a culprit-absent lineup.

There are other circumstances; for instance, photographs of a police surveillance video may bear a resemblance to other witnesses selecting their suspect looks like. This is simulating confederate-culprit researchers select like it, we recommend paragraph which the most-chosen rate can with the most-curious.

The issue of hypotheses about the selection rate (e.g., the effects it is important to important. In g fillers. One is the culprit (as give description from). Another is to the res-
rethought the procedure. One of the issues present and absent in the presence of the question—whether the focus is on the present rather than the future of the statement feedback—has been the question whether and how the results concern present and not simultaneous time. This concerns the focus on the present and the future.

In the same way that is typically, this is a known as unknowingly glossed in the literature. How it is related or merely fits the eyewitness. In this way, the most-chosen innocent person was selected by witnesses. This method can be coupled with a reporting of the rate at which the most-similar foil.

The issue of how to select fillers confronts virtually every eyewitness identification experiment. Obviously, if the experiment itself is testing hypotheses about the best way to select fillers, the hypotheses themselves dictate the selection method. But if the experiment is testing some other question (e.g., the effects of pre-lineup instructions or the effects of exposure duration), it is important to remember that the method of selecting fillers is still important. In general, researchers have focused on two methods for selecting fillers. One is to select fillers who fit the general verbal description of the culprit (as given by people who are shown the culprit and then asked to give a description from memory), which is called the match-to-description method. Another is to select fillers who show some level of similarity to the culprit (called the resemblance-to-culprit method) as rated by pilot study participants. (Note that how this plays out in any given experiment must be determined in part by how the innocent suspect is selected.)
In actual police practices there is jurisdictional variation. Some jurisdictions select their fillers based on the verbal description of the culprit that was given by the eyewitness(es). Other jurisdictions select the fillers based on their similarity to the suspect. The latter practice (similarity-to-suspect method) is actually different from the way fillers are selected in almost any experiment, because the fillers that are selected are more likely to be different if the suspect is innocent than if the suspect is guilty. In other words, the similarity-to-culprit method is not the same as the similarity-to-suspect method. Regardless of how an experimenter decides to select the fillers to be used in the study, it is essential that the method be fully described in any write-up of the study. Furthermore, researchers should usually report some measure of the final lineup product. Often, this measure is functional size (Wells, Leippe, & Ostrom, 1979) or effective size and defendant bias (Malpass, 1981), both of which involve the collection of additional data from “mock witnesses.” However, it should be noted that these measures are totally insensitive to a bias that occurs when the innocent suspect is selected to resemble the culprit, whereas the fillers are selected to fit the description (Wells & Bradfield, 1999). In such a situation, the functional-size, effective-size, and defendant-bias measures will show no lineup bias, and the innocent suspect will actually stand out.

REPORTING AND ANALYZING THE DATA

IDENTIFICATION DATA

Clearly, there are many different approaches to reporting and analyzing data from an eyewitness identification experiment, and researchers must always consider the unique aspects of their experiments in deciding how to approach the data. Nevertheless, there are problems that reoccur frequently. The most common problem is one of underreporting the data. Too often, for example, researchers will collapse across types of accuracy (e.g., identifications of the culprit in a culprit-present lineup and correct rejections from a culprit-absent lineup) and types of inaccuracy (e.g., identifications of the innocent suspect, incorrect rejections when the lineup included the culprit, and filler identifications in both culprit-present and culprit-absent lineups).

Unfortunately, collapsing across both types of accuracy and both types of inaccuracy masks a lot of interesting and important information. We believe that any eyewitness identification experiment that includes culprit-present and culprit-absent lineups should minimally report the descriptive statistics that are characterized by Table 13.1. This kind of table is useful because it exhausts the possible responses that witnesses can make in response to each lineup type. Of course, this should be done for each condition of the experiment and here we have displayed two conditions—X and Y. Note that reporting both the percentages and the frequencies is important, especially if the conditions do not have equal sample sizes. In some cases, it can also be
useful to report the rate at which each individual filler in the lineup is identified rather than collapsing all fillers into one total.

The format of data presentation in Table 13.1 is based on a presumption that the design of the experiment used a single a priori innocent suspect rather than using the method of rotating the innocent suspect across all lineup members (the counterbalancing procedure discussed in the previous section). This is apparent from the fact that the innocent suspect in Condition Y received 30% of the choices whereas the five fillers in total received only 20% (an average rate of 4%). If the rotation method had been used, the innocent suspect rate of identification would be exactly the same as the filler rate divided by the number of fillers.

The format of Table 13.1 also permits a quick assessment of whether eyewitness performance is running above levels expected by chance. One way to define chance is to compare the rate of correct rejections (no identification in culprit-absent conditions) to incorrect rejections (no identifications in culprit-present conditions). Because the frequencies are reported in Table 13.1, anyone can quickly calculate a chi-square value to determine whether the rates are significantly different. The format of Table 13.1 also permits a quick assessment of diagnosticity.

Although space does not permit us to give a full treatment of the diagnosticity statistic, the diagnosticity of a suspect identification is the ratio of hits (identifications of the culprit) to false identifications (identifications of the innocent suspect). In the case of Condition X in Table 13.1, the diagnosticity ratio is 6.5, and in the case of Condition Y, the diagnosticity ratio is 2.5. In effect, the diagnosticity ratio for identifications of the suspect can be interpreted as how much more likely an eyewitness is to identify the guilty party from a culprit-present lineup than to identify an innocent suspect from a culprit-absent lineup. A diagnosticity ratio of 1.0 indicates no diagnosticity, and higher numbers indicate increasing levels of diagnosticity.
The diagnosticity ratio can be also computed for fillers and for no-identification decisions. Diagnosticity ratios are quite important for any analyses that attempt to use Bayesian statistics to calculate probabilities of error under different possible base rates for whether the lineup includes the culprit. Detailed treatments of eyewitness identification diagnosticity and Bayesian probabilities can be found in the eyewitness identification literature (Wells & Lindsay, 1980; Wells & Olson, 2002; Wells & Turtle, 1986). The article by Wells and Olson also describes how to test whether a diagnosticity ratio is significantly different from 1.0 or whether one diagnosticity ratio is significantly greater than another diagnosticity ratio.

Commonly, researchers first want to know whether the manipulation (e.g., conditions X vs. Y in Table 13.1) had any significant effect on eyewitness identification performance. Analyses of variance are generally not appropriate for data of this form (i.e., categorical data of a dichotomous form). An omnibus (overall) test of statistical significance is readily available, however, using a hierarchical log-linear analysis, which is contained in SPSS and other statistical packages. Individual contrasts (e.g., is the 10% rate of innocent suspect identifications in condition X statistically different from the 30% rate of innocent suspect identifications in condition Y?) are also easily conducted using the chi-square statistic.

**CONFIDENCE DATA**

Commonly, researchers are interested in using confidence measures to assess the relation between confidence and accuracy. The point-biserial correlation has been the most common method for doing this in the eyewitness identification literature, but numerous researchers have questioned the appropriateness of the point-biserial correlation and have instead argued for measures of calibration and resolution (e.g., Brewer, Keast, & Rishworth, 2002; Brewer & Wells, 2006; Juslin, Olson, & Winman, 1996). One problem is that calibration and resolution statistics require either extremely large sample sizes or a large number of repeated measures per participant. Note, however, that calibration can be calculated properly only if the confidence scale is categorical from 0% to 100% (i.e., the commonly used 7-point scale cannot be used). Further complicating the traditional practice of reporting the point-biserial correlation between confidence and accuracy is the fact that the point-biserial correlation...
correlation is sensitive to base rates (McGrath & Meyer, 2006). In the current context, that means that the base rate for accuracy affects the point-biserial correlation.

Regardless of the summary statistic used to express the relation between eyewitness identification confidence and accuracy, we recommend that researchers also consider reporting confidence level using accuracy/frequency arrays. Table 13.2 gives an example of such an array. This array is extremely useful because it permits readers to calculate the point-biserial correlation, or $d$, or any other statistic. Importantly, it also permits those who are doing meta-analyses to easily collapse across studies. And, it allows researchers to explore the idea that extreme confidence (e.g., 90% or more confident) might be populated almost exclusively by accurate witnesses (in this case 89% of witnesses with confidence of 90% or better are accurate) even when the overall distribution shows a modest or weak relation.

It is now well documented that the relation between eyewitness identification accuracy and confidence is consistently moderated by whether the witnesses are choosers (i.e., made an identification) or non-choosers (e.g., see Sporer, Penrod, Read, & Cutler, 1995). Hence, it is not recommended that researchers collapse over both types of accuracy (accurate identifications and correct rejections) and both types of inaccuracy (mistaken identifications and incorrect rejections) when reporting confidence-accuracy relations, because this can hide the fact that the relation is stronger among choosers (accurate vs. inaccurate identifiers) than it is among non-choosers (correct rejecters vs. incorrect rejecters).

Finally, we note that researchers should consider the fact that the confidence-accuracy relation in eyewitness identification might also vary as a function of whether the identification of fillers in the culprit-present lineup are included among the choosers in calculating this relation. From a perspective of the ecology of the forensic situation, in actual cases we would not normally be interested in the question of whether eyewitness identification confidence helps to discriminate between identifications of a guilty suspect and identifications of fillers. Using the assumption of the single-suspect model for lineups (one person is the suspect and the remainder are known to be innocent fillers), it is already known in an actual case whether the identification of a filler was mistaken—clearly it was. Instead, eyewitness identification confidence is useful to the extent that it helps discriminate between accurate identifications (i.e., when the culprit is present) and mistaken identifications of an innocent suspect (i.e., when the culprit is absent).

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accurate</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>10</td>
<td>34</td>
<td>21</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Inaccurate</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>12</td>
<td>15</td>
<td>15</td>
<td>33</td>
<td>17</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 13.2
Frequency Array for Confidence Level by Accuracy
In some cases, it might be informative to report the average confidence with which every lineup member was chosen; the suspect in the target-present lineup, each individual filler in the target-present lineup, and each individual member of the target-absent lineup. If standard deviations are also reported with each of these 12 means, it would be easy for the interested reader to calculate the confidence-accuracy correlation under any number of assumptions (e.g., if the most-selected filler was the innocent suspect versus the average filler).

Readers should note that many of the points we have made here about confidence apply as well to other continuous-variable “postdictors” of eyewitness identification accuracy, such as identification decision latencies. Again, however, the approach that a researcher takes can vary depending on the nature of the questions being asked. In the case of decision latencies, for example, Dunning and Perretta (2002) were interested in finding the number of seconds that represented the optimum separation of accurate and inaccurate witnesses. Hence, for their purposes it was useful to find the point along the decision latency time line that yielded the maximum chi-square value separating the accurate from the inaccurate witnesses.

ARCHIVAL RESEARCH AND FIELD EXPERIMENTS

Although our focus in this chapter has been on the design of experimental studies, it is important to note that other methods have been used to examine eyewitness performance. Here we consider two forms of research involving actual cases: (1) archival studies, which use police or prosecutor files to examine past cases or collect new data in police settings but do not involve any experimental methods, and (2) field experiments, which use experimental methods in connection with actual cases. A handful of studies of eyewitness identification performance in actual cases have made use of archival records. The largest and best of these studies have been conducted in the United Kingdom (e.g., Pike, Brace, & Kynan, 2002; Slater, 1994; Valentine, Pickering, & Darling, 2003; Wright & McDaid, 1996) and have involved nearly 15,000 eyewitnesses. Archival studies in North America have been much smaller in scale (e.g., Behrman & Davey, 2001; Behrman & Richards, 2005; Tollestrup, Turtle, & Yuille, 1994).

By field experiments, we mean experiments using eyewitnesses to actual crimes, usually conducted in conjunction with a police department. We are not using the term field experiment to refer to an experiment involving a simulated crime that happens to be conducted in the “field” (e.g., on the street or in a convenience store). (In the latter case, we consider these to be lab experiments that happen to step outside of the traditional lab and set up a lab in some other setting.) A superb example of a field experiment is one conducted by Wright and Skagerberg (2007), in which actual eyewitnesses to serious crimes were asked critical questions about their lineup identification either before or after being told whether the person they identified was the suspect in the lineup. If the issue involved the highly confident suspect, becaus...
Eyezttitness Identification Research 249

...confidence with the target-present confidence of the individual who also reported the suspect to the reader to the reader of assumption versus the assumptions of eye-...tlatencies. Depending on the number of accurate and incorrect point...in chi-square analysis...S

...E...t 1...t 1...s and...or archival and field experiments, however, the identification of a suspect might or might not be an accurate identification. In all of the DNA exoneration cases involving mistaken identification, for example, the identification that was made was the identification of a suspect, albeit an innocent one.

One way to partially get around this problem in archival studies and field experiments is to note that the identification of a filler is always an error whereas the identification of a suspect would sometimes be an error and sometimes not. In the UK studies, approximately one in three positive identifications made by witnesses was a foil identification. However, even this strategy for determining errors can be problematic in archival studies and field experiments due to poor record keeping. For example, Tollestrup et al. (1994) observed that the Royal Canadian Mounted Police did not distinguish between misidentifications of foils and rejections of lineups—both were recorded as a failure to identify anyone. Behrman and Davey (2001) reported the same problem in their study, and it is clear from footnote 3 in an appendix to the Chicago experiment that foil identifications were not recorded because they were considered “tentative,” at least in conditions where the administrator knew the identity of the suspect.

Of course, even without knowing whether the identification of the suspect is accurate, one can assume that such an identification is more likely to be accurate than is the identification of a filler. In this sense, it could be argued that variables that increase identifications of suspects or decrease identifications of fillers are promoting accuracy, whereas variables that decrease identifications of suspects or increase identifications of fillers are promoting inaccuracy. Still, great caution must be exercised in using this logic. Consider, for instance, the myriad ways that one might increase identifications of suspects and decrease identifications of fillers in a field experiment. For example, comparing a condition in which witnesses are kept blind as to which person is the suspect with a condition in which witnesses are told prior to their identification decision which person is the suspect and which ones are merely fillers. The latter would serve to depress the identification of fillers and...
increase identifications of the suspect. Surely we would not accept this as
evidence that witnesses should be told which person is the suspect and which
are fillers.

Alternatively, consider what happens if the fillers in the lineup are selected
specifically to be implausible lineup members (e.g., do not fit the witness’s
description of the culprit). That would also serve to depress filler identifica-
tions and increase identifications of the suspect. But does that promote
accuracy? As a final example, if the depression of filler identifications is a
good outcome, then why use fillers at all? Why not make every lineup a show-
up instead? The point here is that it is not automatically true that variables
that decrease filler identifications or increase suspect identifications in field
experiments necessarily promote accuracy (Wells, 2008).

Similar problems can emerge when examining the confidence-accuracy
relation in archival and field experiments. Again, the idea is that filler
identifications are definitely inaccurate whereas identifications of suspects
have at least some (perhaps the most) accurate identifications in the mix.
Hence, the relation between eyewitness identification confidence and the
filler versus suspect identification decisions can be a proxy for the confidence-
accuracy relation in field experiments. Caution must be exercised here as
well, because there is compelling lab-based evidence that non-blind lineup
administrators (who are the ones soliciting the confidence statement) influ-
ence the confidence of the witness based on the lineup administrators’
knowledge of whether the witness identified a filler or the suspect (e.g.,
see Garrio & Brimacombe, 2001). Hence, in any jurisdiction in which the
lineup administrator is the case detective or someone who knows which
person is the suspect and which ones are fillers (which is the case in most
jurisdictions in the United States), the relation between witness confidence
and whether the witness picked a filler instead of the suspect is likely to be
artificially inflated by the lineup administrators’ influence on the witness.

In general, field experiments that are designed to examine the accuracy
or confidence of witnesses should routinely use double-blind procedures
(Wells, 2008). This means that neither the case detective nor any other
individual who knows which members are fillers and which is the suspect
should be the one who administers the lineup.

Finally, we note that field experiments also differ from lab experiments
in numerous ways that increase the need for quite large sample sizes. In lab
experiments, for example, it is possible to give all witnesses the same type
and quality of view, the same duration of exposure, the same delay between
witnessing and testing, the same fillers, and approximately the same levels
of anxiety, fear, arousal, and so on. In a field experiment, in contrast, these and
numerous other variables float freely and vary widely from one witness to
the next. Careful measures of these free-floating variables can help at the
level of data analysis if they are statistically controlled, but in general, they
contribute noise to the data that results in the need for larger sample sizes
than are needed in the typical lab experiment.
GENERALIZING RESEARCH FINDINGS

Our emphasis in this chapter has been on experimental methods. We have noted a number of the strengths of experimental methods and some of the advantages of laboratory and staged-event experiments over archival studies and field experiments involving actual cases. Our emphasis naturally raises the question of whether the results from laboratory and staged-event experiments generalize to real-world witnesses and what evidence we might use to address that question. Although space does not permit us to undertake a full review of relevant research findings, we do wish to highlight some of the ways in which the question can be addressed and note a few example findings.

DO EXPERIMENTAL STUDIES AND STUDIES WITH ACTUAL WITNESSES YIELD CONVERGENT FINDINGS?

A number of archival researchers have endeavored to test whether findings revealed in experimental studies are replicated with actual witnesses. For example, laboratory studies indicate that the probability of identifying a perpetrator is reduced by the presence of a weapon (e.g., E. F. Loftus, Loftus, & Messo, 1987; Maass & Kohnken, 1989), and Steblay (1992) reported a reliable impairment from the presence of a weapon in a meta-analysis of 19 studies. In their archival study, Tollestrup et al. (1994) found that the presence of a weapon reduces the likelihood of the suspect's being identified: 47% of robbery suspects in cases involving a weapon were identified, as compared with 71% of suspects in robberies without a weapon. In contrast, Behrman and Davey (2001) found similar rates of identification of police suspects when they compared 240 witnesses to crimes in which a weapon was used with 49 witnesses in crimes without weapons. In their study of 640 witnesses, Valentine, Pickering, and Darling (2003) found that the presence of a weapon had no effect on the likelihood of identifying a suspect, but witnesses were more likely to identify a foil if a weapon was not present. Similar comparisons of experimental and archival findings involving other variables have been made by Valentine et al. along with others, and it is not uncommon to see that the findings do not converge. Does this mean the experimental findings are wrong?

We believe it is entirely too early to reach any conclusions from comparisons of this kind. Certainly there is little doubt that a great many variables have been shown to reliably affect witness performance using laboratory and staged-event experimental methods (see, e.g., the review by Wells, Memon, & Penrod, 2006). However, archival and field-experiment data suffer from a number of deficiencies. As we have noted previously, archival studies and field experiments generally cannot authoritatively establish whether witnesses have made correct identifications. They rely on suspect choices as a primary variable, and because suspect choices can involve a mixture of correct and incorrect identifications, researchers should not be too surprised
to observe a different pattern of results when they compare such studies with results from experiments in which accuracy is known.

Another important difference between experimental studies and field or archival studies is that field and archival studies cannot easily isolate variables, because those variables are correlated with other variables, many of which might not have even been measured. Consider, for example, a field study that involved numerous witnesses to a shooting in which it was found that those who were most stressed by the shooting were the most accurate in reporting details of the event (Tollestrup et al., 1994). This finding seemingly goes against the general consensus of the experimental literature on stress and accuracy. However, a closer examination of the data also shows that witnesses who were physically closest to the shooting were the most stressed (which makes sense), and those closest to the event had the better view and were more accurate. These are called multi-collinearity problems, and they are rampant in field and archival studies, which makes it difficult to reach conclusions about the role of individual variables.

There are certainly other features of field data that can muddy the results from such studies. Consider, for example, that in field studies it may be difficult to determine whether a weapon was actually visible to a witness for how long. In addition, not all witnesses are necessarily tested when the police have a suspect. Witnesses who report a poor memory (those who saw a weapon?) may indicate that they do not really remember what the perpetrator looked like and might, therefore, not be shown a suspect. Furthermore, one might ask about case “selection” effects. Are the police differentially selecting weapon versus non-weapon cases for more intensive investigation (we hope so)? Does more intensive investigation of weapons cases yield a higher proportion of guilty suspects (and a higher proportion of suspect choices) as compared with non-weapons cases? We do not know the answer to those questions, though they are subject to research.

Do Different Research Methods Produce Different Results?

As we highlighted earlier in this chapter, researchers are confronted with a variety of methodological choices when they construct their research studies, giving rise to the following questions:

- Should a long or short video be used rather than a staged event containing one or several culprits?
- Should the culprit be replaced with a look-alike when making culprit-absent arrays?
- How many fillers should be selected and in what manner?
- Do these choices yield differences in results that might prompt us to be concerned about the generalizability of our findings?

We would point to two types of analysis that address the question.
Shapiro and Penrod (1986) published a meta-analysis of factors influencing facial identification. They drew on results from over 190 individual studies reported in 128 articles on face recognition and eyewitness memory. Overall, the studies contained 960 experimental conditions under which identifications were made and included more than 16,950 participants. Most (80%) of the studies analyzed by Shapiro and Penrod (1986) were laboratory-based studies of facial recognition rather than of more realistic eyewitness simulations, such as videotaped or staged events. They noted that type of study (face recognition vs. eyewitness) accounted for 35% of the variance in performance across experimental conditions. Performance levels were higher in the laboratory studies, which raises the question of whether the results are generalizable across research methods.

Similarly, Lindsay and Harvie (1988) conducted a meta-analysis that compared a sample of face recognition and staged-crime studies. They found that the overall correct identification rate (.64) in 113 face-recognition studies was significantly higher than the rate (.58) in 47 staged-crime studies and that the false alarm rate (.18) in 64 face-recognition studies was significantly lower than two types of mistaken identifications in staged-crime studies—those made from 47 target-present arrays (.29) and 12 target-absent arrays (.41).

Although these comparisons of laboratory-based face recognition and more realistic eyewitness studies initially may seem to raise concerns about external validity, the difference in performance between the two types of studies would not raise such concerns if explained by variables (e.g., retention interval) that generally differ in face-recognition versus eyewitness studies. Shapiro and Penrod (1986) tested whether “study characteristics” might explain the differences in identification accuracy rates. They found that variables associated with attention variables (including knowledge about an upcoming test and mode of presentation) made significant individual contributions. The duration of exposure per face, pose, memory load at study (number of faces studied), and retention interval accounted for differences in performance. Once all those variables were accounted for, the type of study (facial recognition vs. eyewitness) accounted for just 3% of performance variance (rather than 35%).

A second way to address the generalizability issue was addressed in detail by Penrod & Bornstein (2007), who examined meta-analyses that have appeared since Shapiro and Penrod’s (1986) study. These later meta-analyses permit comparisons of performance across research methodologies that vary over studies; these meta-analyses often ask whether methodological variables interact with substantive variables. Penrod and Bornstein show that these meta-analyses consistently demonstrate that whenever the effects of substantive variables vary across research methodologies, they are larger for more realistic procedures. Factors such as lineup presentation (simultaneous vs. sequential), weapon focus, stress, live versus video testing, instructions to witnesses, cross-race identifications, and the like exert larger effects on eyewitness performance when testing conditions more closely match real
witnessing conditions (e.g., staged crime or other eyewitness event and live video stimuli) than when the situation is relatively contrived and controlled. This pattern of results suggests that if there is a generalizability issue, it is that much eyewitness research underestimates the magnitude of effects being studied. Based on research to date, researchers and practitioners can be reasonably confident that experimental findings generalize to real eyewitness situations.

CONCLUDING REMARKS

In this chapter, we have described a number of methodological issues that confront every researcher who designs and conducts an eyewitness identification study. In many cases, researchers fail to think these issues through fully and end up learning the hard way by having their manuscripts rejected on methodological grounds or discovering the problem later and having to redo the study.

We think this chapter might be especially useful for novice eyewitness researchers. Sometimes we are surprised, however, to find even experienced eyewitness researchers failing to consider some of these issues, and neither of the authors of this chapter can claim to be fully exempt from having made a bad call on one of these methodological issues. No study is perfect, and research involves a constant process of improvement.

We also want to emphasize that with only a few exceptions, our recommendations in this chapter are not universal rules applicable to all studies. Researchers need to consider the issues we have raised in the context of their hypotheses and goals. We gave examples of some instances of this sort, such as when it is acceptable to use only target-present or only target-absent lineups. However, when the question being tested is purely theoretical, some of the ecological validity considerations can become moot.

Finally, we encourage would-be eyewitness identification researchers to read the eyewitness identification literature carefully to gain a sense of other methodological issues that are often unique to a given study. But, we caution researchers not to assume that a study is devoid of all possible methodological problems merely because it was published. The standards for methodological soundness in eyewitness identification research and completeness of reporting are higher today than they were only a few years ago, and this trend will undoubtedly continue.

REFERENCES


false positives, it is that
completeness
some eyewitness
lineups. Applied Cognitive Psychology, 17, 969-993.