

Field Experiments on Eyewitness Identification: Towards a Better Understanding of Pitfalls and Prospects

Gary L. Wells

Published online: 3 July 2007

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

Abstract The Illinois pilot program on lineup procedures has helped sharpen the focus on the types of controls that are needed in eyewitness field experiments and the limits that exist for interpreting outcome measures (rates of suspect and filler identifications). A widely-known limitation of field experiments is that, unlike simulated crime experiments, the guilt or innocence of the suspects is not easily known independently of the behavior of the eyewitnesses. Less well appreciated is that the rate of identification of lineup fillers, although clearly errors, can be a misleading measure if the filler identification rate is used to assess which of two or more lineup procedures is the better procedure. Several examples are used to illustrate that there are clearly improper procedures that would yield fewer identifications of fillers than would their proper counterparts. For example, biased lineup structure (e.g., using poorly matched fillers) as well as suggestive lineup procedures (that can result from non-blind administration of lineups) would reduce filler identification errors compared to unbiased and non-suggestive procedures. Hence, under many circumstances filler identification rates can be misleading indicators of preferred methods. Comparisons of lineup procedures in future field experiments will not be easily accepted in the absence of double-blind administration methods in all conditions plus true random assignment to conditions.

Keywords Eyewitness · Lineups · Field experiments

Eyewitness Identification Field Experiments: Blind Controls, Outcome Measures, and Interpretations

Creating events that are witnessed by unsuspecting people has been a staple methodology for eyewitness scientists. Because researchers created the events, they know the identity of the “perpetrator” with certainty and can manipulate lineup variables systematically to study patterns of identification error. Over time, the emergence of recurring patterns can lead to conclusions about variables that lead to higher and lower rates of error and a better understanding of the underlying processes in eyewitness identification. Not surprisingly, critics contend that these eyewitnesses are not “real” eyewitnesses in the sense that the crime was not real and the consequences of error in real cases will make witnesses too cautious to make these errors. Researchers counter these critics by pointing to DNA-based exonerations that are dominated by cases of mistaken identification and note that the serious consequences (including death sentences) were clearly present in all those cases. Critics counter that there are only 200 such cases that have been proven. Researchers counter that only a small fraction of cases can be discovered with DNA testing, and so on. Clearly, this type of back and forth might be resolvable if well-controlled field experiments with actual eyewitnesses were conducted. It was in this spirit that the Illinois State Legislature mandated that a study be conducted with actual eyewitnesses to test whether the sequential and double-blind lineup procedures are better than simultaneous and non-blind lineup procedures¹.

¹ Sequential lineups show the eyewitness only one lineup member at a time whereas simultaneous lineups show the eyewitness all lineup members at once. Double-blind lineup procedures are administered by someone who does not know which lineup members are fillers (known innocents) and which one is the suspect (who might or might not be the culprit).

G. L. Wells (✉)
Psychology Department, Iowa State University, Ames, IA
50011, USA
e-mail: glwells@iastate.edu

The study was designed and managed by the Chicago Police Department General Counsel Sherry Mecklenburg, who also wrote the report (Mecklenburg 2006a, hereafter called the Mecklenburg Report). There are a number of critiques of the study (e.g., Sherman 2006; O'Toole 2006) and it is not my purpose to spend these pages tearing down a study that was likely based on good intentions. Nevertheless, there are important lessons to be learned from the Illinois study concerning the methods, measures, and interpretations of field experiments on lineups. Learning these lessons now can help prevent various problems in future field experiments on lineups. Accordingly, reference to problems in the Illinois study in the current article are not meant to look backward in an disapproving manner, but are instead in the spirit of looking forward to new, improved field studies.

The Main Problem

As noted by Schacter et al. (2007, this issue), the Illinois study contains a central and serious confound. Specifically, the sequential lineups were always conducted using double-blind procedures and the simultaneous lineups were always conducted using non-blind procedures. Hence, we cannot be certain whether the results (fewer filler identifications and more suspect identifications for the non-blind simultaneous than for the double-blind sequential) are attributable to the sequential versus simultaneous difference or to the double-blind versus non-blind difference. The interpretation matters. In fact, since the time that the double-blind lineup idea was first advocated (Wells 1988), the argument has always been that a non-blind lineup administrator can inadvertently cue eyewitnesses to avoid selecting fillers from lineups and shape them toward identifying the suspect. Hence, if this is the reason that the non-blind simultaneous lineups produced fewer filler identifications and more suspect identifications than the double-blind sequential lineups, then the results constitute a type of proof that lineups should be conducted using double-blind methods.

Clearly, the solution is to compare sequential lineups to simultaneous lineups when both are conducted using double-blind methods. But even this experimental design is missing an important feature of laboratory-based simulated crimes, namely the guilt status of the suspect. Specifically, we cannot be sure that the identification of a suspect is an accurate identification. Suspects are merely individuals who are suspected of the crime. In every DNA exoneration case involving mistaken identification, the eyewitness identified the suspect but the suspect was not the perpetrator. Hence rates of identifying the suspect cannot themselves be used to determine which method produces the fewest errors. But, what about filler identification rates as an outcome variable?

Do Rates of Filler Identification Reveal the Best Procedure?

Theoretically, a filler identification rate is a special outcome variable. Lineup fillers are known-innocents who are placed in the lineup simply to fill it out and make it fair. Therefore, unlike suspect identifications, filler identifications are clearly errors regardless whether they occur in a lab-based simulated crime studies or in real cases. Clearly, the consequences of identifying an innocent suspect (possible charges, possible conviction) are severe compared to the error of identifying a filler (“I’m sorry Ms. Jones, you identified Officer Krupke”). But can rates of filler identification be considered a proxy measure for risk to innocent suspects? And, if so, would not a procedure that produces low filler identification rates always be preferred to one that produces a higher filler identification rate?

Under some conditions, filler identification rates can be considered a proxy measure for the risk to innocent suspects arising from a given lineup procedure. Hence, under special circumstances, the lineup procedure that yields the lowest rate of filler identifications might be considered the preferred procedure. Unfortunately, there are many conditions in which filler identification rates can lead to totally inappropriate conclusions. The most obvious examples arise when considering what would happen if a lineup procedure that was clearly biased against the suspect was compared to an unbiased procedure. Suppose, for instance, that prior to administering the lineup the administrator simply told the eyewitness which members are merely fillers. Comparing this biased procedure against one in which the witness was not given this information would show that the biased procedure produces a lower filler identification rate. If the better procedure was presumed to be the one producing the lowest filler identification rate, then the biased procedure would have to be preferred. This example is absurd because no one would ever set out to conduct a test on such an obviously biased procedure. Suppose, however, the question involved comparing lineups in which the fillers that are used are carefully matched to the witnesses’ descriptions of the culprit versus one in which the fillers do not match the description. Here, we would almost certainly find that the filler identification rate would be much lower when the fillers do not fit the description, a reliable finding in lab studies (e.g., Wells et al. 1993) and one that makes strong common sense. Again, we would be misled if we assumed that lower filler rates reveal the better procedure.

Although the above examples seem obvious, it might be less obvious if we considered what would happen if a double-blind lineup procedure were compared to a

non-blind lineup procedure. A double-blind lineup procedure is one in which the lineup administrator is someone who does not know which lineup member is the suspect and which are merely fillers (Wells 1988). This contrasts with the standard procedure in which the case detective administers the lineup. The case detective will always know which person is the suspect and which are fillers. Psychological science has made a very strong case for the contention that testers influence the person they test in ways that are consistent with the testers' expectations, assumptions, hopes, and so on (e.g., see meta-analysis by Harris and Rosenthal 1985). This phenomenon, often referred to as the experimenter expectancy effect or the interpersonal expectancy effect, does not require awareness or intent on the part of the tester. It is the fundamental reason why double-blind testing is the standard for scientific experiments. These influences can be verbal (e.g., "Did you get a good look at number three?") or non-verbal (e.g., smiles, frowns, nods) and need not be blatant in order to have powerful effects. If the eyewitness appears to be ready to select a filler from a lineup, for instance, the lineup administrator might say "Take your time... be sure that you have looked at all the photos." Even if there were an outside observer, the observer might not recognize that such a comment has the effect of leading the eyewitness away from the filler. This is especially true if the eyewitness believes that the lineup administrator has special knowledge of which lineup member is the suspect and which are fillers, which would always be the case for a non-blind lineup procedure.

The interpersonal expectancy effect makes it easy to see how a double-blind procedure might yield a higher rate of filler identifications than would a non-blind procedure. But, this is really no different than finding that a research experiment comes out "better" (i.e., more in line with the experimenters' hopes, expectations, prior beliefs) if it is conducted in a non-blind manner than if it is conducted using double-blind procedures. Both the lineup and the experiment are improper tests of their respective hypotheses when conducted using non-blind procedures. Showing that a non-blind procedure produces "better" results than double-blind procedures is not an argument for preferring non-blind procedures. If anything, getting what appears to be better results using non-blind procedures is evidence that the procedures should be conducted with double-blind methods.

Because a field experiment on eyewitness identification procedures relies rather heavily on filler identification rates as the only unambiguous error rate, it would never be acceptable to test one procedure (e.g., sequential) that used double-blind techniques and compare the results to another procedure (e.g., simultaneous) that used non-blind techniques.

Other Problems with Non-Blind Lineup Procedures

Compared to double-blind lineup procedures, non-blind lineup procedures are expected not only to suppress filler identification rates and raise rates of identifying the suspects, but also to influence the confidence of the eyewitness in a selective fashion. Lab experiments have shown extremely strong and consistent effects on eyewitness confidence when the lineup administrator reacts to the identification made by the eyewitness. The reaction "Good, you identified the suspect" leads eyewitnesses to believe that they were certain all along and raises their confidence to high levels even if they identified an innocent person (Bradfield et al. 2002; Dixon and Memon 2005; Douglass and McQuiston-Surrett 2006; Hafstad et al. 2004; Neuschatz et al. 2005; Semmler and Brewer 2006; Semmler et al. 2004; Wells and Bradfield 1998; Wells and Bradfield 1999; Wells et al. 2003; see meta-analysis by Douglass and Steblay 2006). Telling eyewitnesses that they identified a filler has an effect in the opposite direction. Importantly, a recent field experiment with actual eyewitnesses to serious crimes shows that this same phenomenon occurs with real eyewitnesses: as soon as they were told that they identified the suspect, eyewitnesses tended to bolster their identification (e.g., saying they had a good view, that the identification was easy to make) but as soon as they were told that they identified a filler, they began to retreat from their identifications (Wright and Skagerberg 2007).

This phenomenon, known as the post-identification feedback effect, is very important and relevant to any field experiment because it affects the very definition of what constitutes an "identification." In the Mecklenburg Report (Mecklenburg 2006a), for example, it was reported that neither Chicago nor Evanston had any instances of filler identifications for their simultaneous non-blind lineups. This absolute zero rate of filler identifications was very surprising given that other jurisdictions consistently report an average of 20.5% filler identifications in actual lineups (Behrman and Davey 2001; Behrman and Richards 2005; Slater 1994; Valentine et al. 2003; Wright and McDaid 1996; Wright and Skagerberg 2007). How can one account for zero percent filler identifications for Chicago and Evanston but 20% for other jurisdictions? Months after the original Mecklenburg Report (Mecklenburg 2006a) an addendum to the Report was released. An interesting observation appeared in a footnote. Footnote 3 on page 8 says:

"There has been some question raised as to why the filler identification rates in both the Chicago and Evanston simultaneous lineups were reported as zero. Although there were in fact filler identifications in simultaneous lineups in those two jurisdictions, those filler identifications were tentative and therefore

under the coding employed by the two analysts, were not reported as actual filler identifications. Similarly, such tentative identifications would not be considered actual identifications in the criminal justice system” (page 8 footnote 3 of *Mecklenburg, 2006b*).

It remains unclear exactly what the criteria were for not counting these filler identifications from the non-blind simultaneous lineups. Presumably, these eyewitnesses who had identified fillers were thought to be not very confident based on their verbal comments following their identifications and hence were not recorded as filler identifications. But, were the double-blind (and sequential) and the non-blind (and simultaneous) filler identifications on equal footing here? As noted earlier in this article, as soon as an eyewitness learns (through verbal or non-verbal reactions of the lineup administrator) that she or he identified a filler, they begin to express lower confidence in their identification. Did those who identified fillers in the simultaneous lineups express low confidence because the lineup administrators were not blind when the simultaneous lineups were used? The beauty of a double-blind procedure is that the persons administering the lineups do not know whether the witness identified a filler or identified the suspect. Hence, using double-blind procedures the lineup administrators cannot have dismissive reactions to filler identifications and would have to code filler identifications as identifications (because they might have been identifications of the suspect for all the lineup administrators knew). Non-blind lineup administrators, on the other hand, know immediately whether the identifications were of fillers and their verbal or nonverbal reactions could easily make the eyewitnesses express low confidence. In fact, a study by Garrioch and Brimacombe (2001) demonstrated exactly that phenomenon. These researchers manipulated lineup administrator’s beliefs about which person was the suspect and which were fillers. When the lineup administrator was led to believe that the witness picked a filler, the confidence statement obtained by that administrator from the eyewitness was much lower than if the administrator was led to believe that the identification was of the suspect.

The point here is that non-blind administration of lineups can not only reduce the frequency of filler identifications, but also non-blind administration will tend to lower the witness’s certainty in their filler identifications. As a result, non-blind administration will end up not counting many filler identifications that a blind administration would have counted.

General Lessons for Field Experiments on Lineup Procedures

Space does not permit a full treatment here of all of the controls needed for definitive field experiments on

eyewitness identification procedures. Instead, I have focused on one major issue, namely the precarious status of filler identifications. This is an important focus because it is tempting to assume that, because filler identifications are “known errors” in actual lineups, a procedure that produces a lower rate of filler identifications must be the better procedure. But, I have described how filler identification rates can be suppressed by clearly unacceptable procedures and how non-blind lineup administration will suppress filler identifications for inappropriate reasons that make interpretation impossible.

How can a field experiment ever be conducted in a manner that yields interpretable results as to whether one lineup procedure is better than another? Clearly, that can never be done using non-blind procedures. So, one lesson is to make sure that all conditions of any such experiment use double-blind methods. Not only should double-blind methods be used to administer the lineups, but double-blind methods should be used to score any data for which there is room for interpretation. For instance, in assessing the confidence of an eyewitness from their verbal confidence statement, the scorer should not know if the witness identified the suspect or a filler.

Second, the experiment must use true random assignment to conditions. In Chicago, two sites were used: All the double-blind sequential data were obtained from Area 011 of District 4 and all the non-blind simultaneous data were obtained from the remainder of District 4. Clearly, that was not random assignment to procedures. True random assignment for a field lineup experiment means that each eyewitness in the experiment has an equal chance of being tested using one or the other procedure regardless of all other considerations. The idea of true random assignment is that any differences between one procedure and the other procedure are due to chance alone plus the effect of the procedures themselves. Because statistical methods for calculating chance are well developed, the effect of the procedures themselves can be estimated when true random assignment is used.

A third lesson is that the random assignment must be made at a time that prevents the possible introduction of other systematic differences. Consider, for example, a test of the simultaneous versus sequential procedure. If the person who puts together the lineup (i.e., selects fillers) knows a-priori that the lineup will be presented using the simultaneous versus the sequential procedure (which they did in the Illinois study), then this could influence how that lineup is put together (e.g., better fillers on average for one procedure versus the other). The optimal time for random assignment to simultaneous versus sequential is for this to occur after all other decisions (e.g., whom to use as fillers) have already been made.

A fourth lesson is that field experiments must preserve the actual lineups (photos of the live lineup, high quality copies of the photo lineups) for additional analyses. Rates of suspect and filler identifications are highly sensitive to the similarities among lineup members. If these similarities are not equal between the two procedures being tested, then the comparison of procedures is confounded.

A final lesson is that field experiments on eyewitness identification procedures must involve a full collaboration between scientists, police, and prosecutors at all stages of the experimental process. The inclusion of scientists in the initial development of the questions, design, procedure, and measures (rather than just data analysis) is critical to the ultimate acceptability of the results.

References

- Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior, 25*, 475–491.
- Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior, 29*, 279–301.
- Bradfield, A. L., Wells, G. L., & Olson, E. A. (2002). The damaging effect of confirming feedback on the relation between eyewitness certainty and identification accuracy. *Journal of Applied Psychology, 87*, 112–120.
- Dixon, S., & Memon, A. (2005). The effect of post-identification feedback on the recall of crime and perpetrator details. *Applied Cognitive Psychology, 19*, 935–951.
- Douglass, A. B., & McQuiston-Surrett, D. M. (2006). Post-identification feedback: Exploring the effects of sequential photospreads and eyewitnesses' awareness of the identification task. *Applied Cognitive Psychology, 20*, 991–1007.
- Douglass, A. B., & Steblay, N. (2006). Memory distortion in eyewitnesses: A meta-analysis of the post-identification feedback effect. *Applied Cognitive Psychology, 20*, 859–869.
- Garrioch, L., & Brimacombe, C. A. E. (2001). Lineup administrators' expectations: Their impact on eyewitness confidence. *Law and Human Behavior, 25*, 299–315.
- Hafstad, G. S., Memon, A., & Logie, R. (2004). Post-identification feedback, confidence and recollections of witnessing conditions in child witnesses. *Applied Cognitive Psychology, 18*, 901–912.
- Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin, 97*, 363–386.
- Mecklenburg, S. (2006a). Report to the Legislature of the State of Illinois: The Illinois Pilot Program on Sequential Double-blind Identification Procedures.
- Mecklenburg, S. (2006b). Addendum to the Report to the Legislature of the State of Illinois: The Illinois Pilot Program on Sequential Double-blind Identification Procedures.
- Neuschatz, J. S., Preston, E. L., Burkett, A. D., Togli, M. R., Lampinen, J. M., Neuschatz, J. S., Fairless, A. H., Lawson, D. S., Powers, R. A., & Goodsell, C. A. (2005). The effects of post-identification feedback and age on retrospective eyewitness memory. *Applied Cognitive Psychology, 19*, 435–453.
- O'Toole, T. P. (2006). What's the matter with Illinois? How an opportunity was squandered to conduct an important study on eyewitness identification procedures. *Champion, 2006*, 18–23.
- Schacter, D., Dawes, R., Jacoby, L. L., Kahneman, D., Lempert, R., Roediger, H. L., & Rosenthal, R. (2007). Studying eyewitness investigations in the field. *Law and Human Behavior* (this issue).
- Semmler, C., & Brewer, N. (2006). Post-identification feedback effects on face recognition confidence: Evidence for metacognitive influences. *Applied Cognitive Psychology, 20*, 895–916.
- Semmler, C., Brewer, N., & Wells, G. L. (2004). Effects of postidentification feedback on eyewitness identification and nonidentification. *Journal of Applied Psychology, 89*, 334–346.
- Sherman, L. W. (2006). To develop and test: The inventive difference between evaluation and experimentation. *Journal of Experimental Criminology, 2*, 393–406.
- Slater, A. (1994). *Identification paradises: A scientific Evaluation. Police Research Award Scheme*. London: Police Research Group, Home Office.
- Valentine, T., Pickering, A., & Darling, S. (2003). Characteristics of eyewitness identification that predict the outcome of real lineups. *Applied Cognitive Psychology, 17*, 969–993.
- Wells, G. L. (1988). *Eyewitness identification: A system handbook*. Toronto: Carswell Legal Publications.
- Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect:" Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360–376.
- Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitnesses' recollections: Can the postidentification feedback effect be moderated? *Psychological Science, 10*, 138–144.
- Wells, G. L., Olson, E., & Charman, S. (2003). Distorted retrospective eyewitness reports as functions of feedback and delay. *Journal of Experimental Psychology: Applied, 9*, 42–52.
- Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). On the selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*, 835–844.
- Wright, D. B., & McDaid, A. T. (1996). Comparing system and estimator variables using data from real lineups. *Applied Cognitive Psychology, 10*, 75–84.
- Wright, D. B., & Skagerberg, E. M. (2007). Post-identification feedback affects real eyewitnesses. *Psychological Science, 18*, 172–178.