Law Hum Behav
DOI 10.1007/s10979-009-9207-7

ORIGINAL ARTICLE

What We Know Now: The Evanston Illinois Field Lineups

Nancy K. Steblay

Abstract A Freedom of Information Act lawsuit secured 100 eyewitness identification reports from Evanston, Illinois, one of three cities of the Illinois Pilot Program. The files provide empirical evidence regarding three methodological aspects of the Program's comparison of non-blind simultaneous to double-blind sequential lineups. (1) A-priori differences existed between lineup conditions. For example, the simultaneous non-blind lineup condition was more likely to involve witnesses who had already identified the suspect in a previous lineup or who knew the offender (non-stranger identifications), and this condition also entailed shorter delays between event and lineup. (2) Verbatim eyewitness comments were recorded more often in double-blind sequential than in non-blind simultaneous lineup reports (83% vs. 39%). (3) Effective lineup structure was used equally in the two lineup conditions.

Keywords Eyewitness · Lineups · Field study

Law and Human Behavior published a series of six commentaries in 2008 that addressed the Illinois Pilot Program, a year-long field test of police lineups in three Illinois cities (Cutler & Kovera, 2008; Mecklenburg, Bailey, & Larson, 2008a; Ross & Malpass, 2008; Schacter et al., 2008; Steblay, 2008; Wells, 2008). A summary of the pilot program produced two years earlier by the General Counsel to the Chicago Police (Mecklenburg, 2006a, b) generated substantial controversy in scientific and legal communities. Mecklenburg suggested existing Illinois police lineup practice to be superior to procedures developed in eyewitness laboratories. That is, non-blind simultaneous-format lineups produced higher suspect identification rates and lower filler picks (known errors) than did the double-blind sequential-format lineups recommended by many scientists (see e.g., Steblay, Dysart, Fulero, & Lindsay, 2001; Wells, Memon, & Penrod, 2006).[1]

The legislated intent of the Illinois Pilot Program (hereinafter the *Program*) was to determine the efficacy of the double-blind sequential lineup compared to Illinois practice, and the Program was to be "designed to elicit information for comparative evaluation purposes, and…consistent with objective scientific research methodology" (Capital Punishment Reform Study Committee Act, 2003). The ensuing arguments regarding Mecklenburg's (2006a) report focused on interpretation of the data given the design of the study, and "objective scientific methodology" became the linchpin for debate regarding the pilot results. For example, the lead article in the LHB series (Schacter et al., 2008) provided an evaluation of the Illinois study design by a group of distinguished scientists. The seven-member panel issued a judgment that the Program was unreliable as a basis for determining effective eyewitness identification procedures. This conclusion reflected the panel's view that the study had a fundamental confound in its comparison of double-blind sequential lineups with non-blind simultaneous lineups, a flaw that has "devastating consequences for assessing the real-world implications…[and] guaranteed that most

Preliminary data were presented at the American Psychology-Law Society Conference, 2009, San Antonio, TX.

N. K. Steblay (✉)
Department of Psychology, Augsburg College, 2211 Riverside Avenue, Minneapolis, MN 55454, USA
e-mail: steblay@augsburg.edu

[1] This outcome was present in two of the three cities tested, Chicago and Evanston. Joliet, however, produced results in line with laboratory results.

Published online: 23 February 2010

outcomes would be difficult or impossible to interpret." (Schacter et al., 2008, p. 4). In short, the study could not answer the research question as to whether sequential lineup procedures are superior to simultaneous, nor whether double-blind procedures are superior to non-blind. The panel also noted the inherent difficulty of interpreting results from a non-blind condition, a design weakness that others had targeted as well (Steblay, 2006, 2008; Wells, 2006b, 2008).

Mecklenburg, Bailey, and Larson (2008a, b) have rejected criticisms of the Program. Important for this current analysis, they have made a further claim for the value of the Illinois data beyond the contested comparison of double-blind sequential with non-blind simultaneous lineups. That is, Mecklenburg et al. suggest that even when comparisons between the two conditions are set aside, within-condition data still offer meaningful descriptive summaries of lineup outcomes under each lineup protocol: "[T]he Illinois data can readily be analyzed outside the context of the confound simply by eliminating the comparison between the two sets of data" (2008a, p. 23). With this assertion, Mecklenburg et al. have proceeded to compare the Illinois results to both lab studies and extant field data from Hennepin County, Minnesota, and Queens County, New York.

This prompts the question: *What exactly was measured in the Illinois Pilot Program?* Practically stated, how similar were Illinois lineup protocols to those of other jurisdictions or to the laboratory? The purpose of this investigation is to explore the Illinois Pilot Program through files recently made available as a result of a Freedom of Information Act (FOIA) lawsuit. These files clarify aspects of Evanston lineup structure and procedure, that is, local police practice. Two primary analyses were conducted. First, descriptive statistics summarized characteristics of the overall Evanston dataset and of the two lineup conditions: non-blind simultaneous and double-blind sequential. A subset of data that met defined practice also was evaluated (witness first attempts to identify a stranger-offender from single-suspect lineups of size 5–6). Second, a mock witness procedure was conducted in order to evaluate lineup quality.

The FOIA Lawsuit

In 2007, a FOIA lawsuit was filed by the National Association of Criminal Defense Lawyers (2007) against the police departments of the three cities of the Illinois Pilot Program—Chicago, Evanston, and Joliet. The lawsuit claimed that, absent essential information in the Mecklenburg (2006a, b) documents, additional facts from police files are necessary to fully inform the public as to the validity of the Illinois Pilot Program results and the report ensuing from that study.

The lawsuit requested information approximate to that required for scientific peer review—a full detailed description of the study design, research protocol, and analyses. The underlying rationale for a scientific review process was described in the lawsuit—that scientific "reviewers will help to identify strengths and weaknesses of the study, question the appropriateness of analyses and interpretations, call for additional analyses as necessary, tamp down overzealous claims, and thereby help to better align conclusions with existing data" (Steblay, 2007a, p. 4). As noted by LHB editors Cutler and Kovera (2008), a peer review process would allow "… a better understanding of what we are arguing about" (p. 2). Peer review also promotes a thoroughness in reporting that allows independent researchers to replicate the study in their own labs or jurisdictions. From a practical standpoint, the relevance of the Illinois results for another jurisdiction will depend on how closely its practices mirror those in Illinois; precision about Illinois police procedure is essential. More broadly, clarification of the Illinois Program is germane to the continuing debate about its relevance for practice, policy, and legal proceedings (see e.g., Kings County Supreme Court, 2008).

The Evanston Illinois Police Department cooperated with the FOIA request, providing case files from 100 field identification tests. For the purposes of this article, two primary categories of previously unavailable information were particularly useful for analysis: witness and suspect identification histories, and makeup of the lineup (the lineup photos).

What Was Measured in the Illinois Pilot Program?

It is imperative to be mindful of existing ambiguities in the Illinois Program data and of the likelihood of resolving these difficulties with the Evanston files. The design confound highlighted by Schacter et al. (2008) cannot be untangled after-the-fact, nor can the direct impact of non-blind lineup administrators upon eyewitness decisions and investigator recording practices be determined. These challenges to interpretation of the Illinois Program must not be downplayed, but unfortunately also cannot be resolved with data garnered from the FOIA lawsuit.

On the other hand, a provocative gap in the Mecklenburg report can be addressed with the FOIA data: the absence of information about each eyewitness's identification history. These details are particularly relevant, given that Illinois police sometimes conduct a physical lineup after the witness has identified a suspect from a photo array. Procedures of repeated identification will likely raise aggregate identification rates from first to final lineups (Behrman & Davey, 2001). A potential problem for understanding the Illinois Program is an underlying reporting issue: inclusion of both first and second lineup decisions from a single eyewitness renders non-independent observations (double-counting). At the least, it is germane to determine how repeated lineups are represented in the aggregate data. Were both first and second identification decisions recorded? Or, if reported only once, which lineup was counted—photo or live, first or second? There may be a collateral impact of the counting method if second identification attempts are included, in that most witnesses who initially fail to identify the suspect from a show-up or early photo array will not advance to the live lineup. Therefore, witnesses brought to live lineups are more apt to produce suspect identifications. Knowledge of this chain of identification events—whether a witness has viewed the suspect in a prior array or show-up—can clarify the study's comparability to other jurisdictions and datasets. Repeated identifications prohibit an apples-to-apples comparison with field data in which multiple identifications were not used (e.g., in Hennepin County Minnesota; Klobuchar, Steblay, & Caligiuri, 2006), or to lab studies that did not employ multiple identification tasks. An additional very serious problem with repeated identification lineups is how this practice affects the veracity of eyewitness memory, a point discussed in more detail later in this article.

A related point of witness history is the relationship between witness and offender, specifically whether they were strangers or acquaintances at the time of the crime. Identifications of known perpetrators will increase suspect identification rates and decrease filler selection rates (Klobuchar et al., 2006). For the purposes of this article, a lineup task in which the witness is asked to attest to the identity of an offender known prior to the crime will be called a *verification lineup*. A lineup presented to a witness who has seen the suspect in a previous identification task will be referred to as a *confirmatory lineup*. The Evanston case files can empirically inform these points as is essential for understanding both descriptive and comparative analyses within the Illinois Program results.

An additional unknown aspect of the Illinois data is the quality of the lineups (Ross & Malpass, 2008; Steblay, 2006). Sound lineup practice is a function of both lineup construction and administration. Seemingly desirable outcomes of high suspect identification rates and low filler selection rates may not be due wholly to good lineup procedure, but in part to poor construction of the lineups. If fillers are implausible, lineup size truncated, or the lineup biased, the suspect will become the witness's logical pick even without true recognition. The Evanston files include 61 photo lineup arrays, allowing a test of filler quality for a subset of the data through a mock witness procedure (Doob & Kirshenbaum, 1973; Steblay, 2007b; Wells, Luus, & Windschitl, 1994).

Finally, an inherent challenge in field data is the absence of ground truth. The accuracy of the key dependent measure—the witness's decision—is difficult to determine in the field because the suspect may or may not be the perpetrator. Numbers that emerge from field tests cannot be translated as absolute measures of right/wrong eyewitness decisions and thus have enormous potential to be misunderstood. Although the FOIA data cannot directly resolve this difficulty in the Illinois Program, a broader view of Evanston police investigations may shed light on effective means for establishing ground truth in future field tests.

Method

Analysis 1: Evanston Eyewitness Identification Reports

Sample. One hundred eyewitness identification reports were provided by the Evanston Police Department, along with contextual case file information.[2] Thirteen identification reports were eliminated from the analysis because they did not involve a full lineup (these were show-ups or mug-book identifications) or did not report lineup outcome or format. Of the remaining 87 lineups, 36 were non-blind simultaneous lineups, 46 were double-blind sequential lineups, and 5 were non-blind sequential lineups.[3]

Procedure. Each case file was examined at least five times by the author. The purpose of the first review was to organize the lineups chronologically within case, to create a codebook and data file, and to complete a coding sheet for each lineup. A second review focused on ancillary file

[2] The Illinois Report does not state the number of original Evanston files, although it can reasonably be assumed that the same set was received by the two data analysts, Drs. Malpass and Ebbesen. Dr. Malpass (the primary analyst) and Ms. Mecklenburg report that they no longer have case files or data. Dr. Ebbesen provided the information that his analyses were based on 101 identification attempts (personal communication, July 15, 2009). These attempts involved 95 single-suspect lineups, of which 19 were live lineups (10 simultaneous, and 9 sequential) (Mecklenburg, 2006b, Table 7, p. 7). Dr. Ebbesen's numbers are near identical to those of the current (FOIA) study, in which 100 reported identification attempts yield data for 93 single-suspect lineups, with 21 live lineups (12 simultaneous and 9 sequential). Dr. Ebbesen reports 43 simultaneous lineups and 52 sequential lineups, numbers which very closely approximate the FOIA set (at 42 and 51, respectively) *if* one takes into account 5 non-blind *sequential* lineups and 6 lineups reported without format (*simultaneous*?) that are noted in the results of this report. For his analysis, Dr. Malpass filtered the Illinois data to screen out non-stranger identification attempts or lineups for which the level of the familiarity between the witness and the suspect could not be determined, and to eliminate lineups in which identifications were reported as "tentative" (Malpass, 2006a, March, Illinois Project Protocol and Code Book: Sim/Seq Lineup Evaluation, pp. 9–10). The Mecklenburg (2006b) report describes 71 Evanston lineups remaining after the Malpass filtering process.

[3] The five non-blind sequential lineups produced 100% suspect identifications; these lineups were excluded from the analysis. Four of the five were identifications of strangers, one of a known offender.

documents, revising and refining the codebook and data file as required. Three weeks later, the coding sheets were set aside and a fresh sheet completed for each lineup; subsequently the two sheets were compared for coding consistency within and across cases, and discrepancies were resolved. At a fourth step, the lineup photos were examined; in some cases these photos alerted the researcher to necessary refinements in the dataset, as when a specific suspect appeared across cases.

After completion of the previous four steps, three student assistants, blind to the specific purpose of the study, were trained to use the coding system. Each coder then independently evaluated a set of lineups, with the end result that every lineup was reviewed by at least two persons: the author and one independent coder. There were two types of coded information. First, most details in the police reports (such as crime type and witness/suspect demographics) required minimal or no interpretation. Coding disparities on these factors were resolved by returning to the files; a detail overlooked by one or the other coder was found and the data sheet corrected as necessary. This practice resulted in 100% final agreement. A second type of detail stemmed from the police investigation as it was reported over time and across lineups, information that was sometimes incomplete or unclear in individual police reports. Witness and suspect names had been redacted in all the Evanston files, thereby producing challenges in tracking suspect/witness identification histories. When differences between coders surfaced, a third was asked to review the case, and subsequent consensus established based on a consistent system. In three cases involving nine lineups (9% of the 100 Evanston reports), ambiguity on a single point of each file was unresolved even with further discussion and document review. These points of ambiguity do not significantly alter conclusions, but will receive comment in the results section.

Analysis 2: Filler Quality

Sample. Seventy-eight college students and community members participated in the study. The sample was 82% Caucasian, evenly divided between females and males (50% each), and ages ranged from 18 to 58 years, with a mean of 23 years, $SD = 7.5$.

Lineup fairness is not a property of a lineup per se, but rather a result of the interaction between the lineup and the verbal description provided by a specific witness (Wells & Bradfield, 1999). Therefore, in order to use the mock witness protocol, it was necessary to find lineups in which the police report included a witness's prior verbal description of a stranger–perpetrator. From the original set of 61 available photo lineups, 19 were found suitable: seven from NB-SIM lineups, 10 from double-blind sequential lineups, one from a non-blind sequential lineup, and one from a lineup without recorded format. All but one were six-member lineups (the exception was of size five). Remaining lineups were eliminated because: (1) the perpetrator was known by the witness; (2) it was not clear that the eyewitness had provided the description used to construct the lineup; (3) multiple witnesses/lineups were conducted for the same suspect; or (4) photo clarity was poor (poor photocopy).

Procedure. Testing was done by three experimenters who were blind as to which lineup member was the suspect. Experimenters left the participant's cubicle during the rating task. Each participant evaluated all 19 lineups, with presentation order of the lineups counterbalanced across subjects. All lineups were presented in simultaneous format. Participants were instructed as follows:

> In this packet is a series of real police lineups. Connected to each one is a description of the perpetrator that was provided by the real witness to the crime. Some of these descriptions are very brief, others provide more detailed information. The entire description provided by the witness is given, even if some features cannot be seen in the photo. For example, an attribute of the perpetrator's body, not visible in the picture, may be part of the witness's description. Also, just as would a real witness, you can assume that clothing and sometimes appearance will have changed between the witnessed event and the time of the lineup photo. Your task for each lineup is to read the description, view the lineup, and make a choice, as best you can, as to *which of the lineup members you think is the accused* (Wells & Bradfield, 1999). That is, who do you think the suspect is?

Results

Analysis 1: Verification and Confirmatory Lineups

The FOIA lawsuit sought to secure information about the context of each lineup, specifically the identification history for each witness and suspect. In the set of 87 Evanston lineups identified above, 25 verification lineups of known or familiar perpetrators produced 84% suspect identifications and no filler picks.[4] Seven confirmatory lineups (all live) produced 60% suspect identifications and no filler

[4] Verification lineups are meant to bring witness and police together in their understanding of a suspect's identity in instances in which the perpetrator is familiar to the witness. It should be acknowledged that this may not produce 100% suspect identifications—the suspect may in fact not be the known perpetrator ("no, that's not the Big Jim I'm talking about")—but this scenario is unlikely to result in a filler pick.

Table 1 Characteristics of 36 non-blind simultaneous (NB-SIM) and 46 double-blind sequential (DB-SEQ) lineups

	NB-SIM (36) (%)	DB-SEQ (46) (%)	*r*	*p*
Verification of known perpetrator	38.9	21.7	.19	*.09**
Confirmation of previous ID (same witness)	16.7	2.2	.26	*.02*
Unreported outcomes	8.3	0	.22	*.05*
Single suspect, first ID, stranger	50.0	73.9	.25	*.03*
Live lineup	33.3	19.6	.16	.16
Six-person lineup	77.8	78.3	.01	.96
Victim eyewitness	61.1	71.7	.11	.31
Bystander eyewitness	33.3	13.6	.24	*.03*
Witness did not see event	5.6	10.9	.09	.40
Delay (same day identification)	27.8	6.8	.29	*.009*
Delay (next day identification)	19.4	2.3	.28	*.01*
Violent crime	69.4	69.6	.00	.98
Witness race (white)	42.4	58.1	.15	.18
Perp race (white)	2.8	11.4	.16	.14
Same-race	51.5	34.9	.17	.14
Weapon	55.6	35.7	.20	*.07*
Number of perps (1)	70.6	73.2	.03	.79
Description in file	86.1	60.9	.28	*.01*
Lineup in file	91.4	63.0	.29	*.01*
Eyewitness comments in own words, in file	38.9	82.6	.45	*.0001*

* *Z*s based on $n = 82$, with minor exceptions for unreported data; *p*s reported are two-tailed

picks. Suspect identification rates increased across categories: from 37% for first identification of a stranger, to 60% for second identification of a stranger, to 95% for identification of a perpetrator known by name.

Verification and Confirmatory Lineups in Simultaneous Versus Sequential Lineup Conditions. Evanston employed an even–odd case number method of assigning lineups to format condition. Of the 82 lineups that met the inclusion criteria listed above, fewer lineups (36 vs. 46) were documented in the non-blind simultaneous (NB-SIM) than in the double-blind sequential condition (DB-SEQ), yet a larger percentage of NB-SIM lineups involved an offender known to the witness (38.9% vs. 21.7%), $Z_p = 1.70$, $p = .09$, or previously identified by the same witness in an earlier array (16.7% vs. 2.2%), $Z_p = 2.34$, $p = .02$ (see Table 1). The NB-SIM condition also included lineups with unreported outcomes, whereas the sequential condition did not (8.3% vs. 0%), $Z_p = 1.98$, $p = .04$).[5]

The impact of verification and confirmatory lineups on the aggregate numbers of the study was examined through a subset with these lineups removed (Table 2, "Adjusted"). This subset also consists of single-suspect lineups of five to six members. Whereas 73.9% of the DB-SEQ lineups met the criteria of first-identification for stranger–perpetrators, significantly fewer of the NB-SIM lineups met the criteria (50%), $Z_p = 2.23$, $p = .03$. Mecklenburg (2006a; Malpass et al., 2006b, Table 3a) documented Evanston suspect identifications in the NB-SIM lineup condition at 67.7%; the current subset shows 50% suspect IDs, a difference of 17.7 percentage points. Suspect identifications in the DB-SEQ lineup condition were slightly higher (29.4% vs. 25.9%) compared to Mecklenburg's report. In short, the Evanston files provide empirical evidence that the strategy employed for random assignment to the two lineup conditions was not effective and that this design error, along with the presence of verification and confirmatory lineups, produced a disproportionate inflation of suspect identifications in Evanston's NB-SIM lineup condition.

Additional areas of dissimilarity exist between the two lineup format conditions (see Table 1). First, aspects of witness memory encoding and retrieval differ. Bystander-witnesses (at a significant level) and crimes involving weapons (at a marginally significant level) were more common in the NB-SIM condition. Delayed identification attempts were significantly more frequent in the DB-SEQ condition, with only 8.7% occurring within 1 day of the crime event, whereas 47.2% of NB-SIM lineups took place in the same or next day after the crime. Although not statistically significant, same-race identifications were more common in the NB-SIM lineup condition (16% higher than DB-SEQ) as were live lineup presentations (14% more). Second, aspects of the police report differed between conditions, with the NB-SIM lineup files significantly more likely to include the lineup photos and witness

[5] *Z*-tests for proportions (Z_p) are calculated with $n = 82$ (exceptions are noted) and reported with two-tailed *p*-values.

Table 2 Eyewitness decisions (%) in non-blind simultaneous and double-blind sequential lineup conditions

	Non-blind simultaneous			Double-blind sequential		
	Mecklenburg $n = 31$	Evanston $n = 36$	Adjusted $n = 18$	Mecklenburg $n = 27$	Evanston $n = 46$	Adjusted $n = 34$
Suspect IDs	67.7	58.3	50.0	25.9	41.3	29.4
Filler picks	0	0	0	11.1	13.0	14.7
Ruled out		16.7	22.2		17.4	23.5
No Choice	32.3	16.7	27.8	63.0	28.3	32.4
Unreported	0	8.3				

Mecklenburg = filtered for known offenders (from 71 to 58; 13 lineups removed)

Evanston = FOIA lineups (82)

Adjusted = Evanston FOIA lineups filtered for confirmatory, verification, and unknown outcomes (87 to 52; 35 lineups removed)

Note: One of the double-blind sequential lineups in this set is arguably not blind—the document claims the detective is unfamiliar with the case, but his/her written record states that the "witness identified a filler photo, not the suspect," a distinction only known by a non-blind administrator. If the numbers above are corrected for this circumstance (double-blind SEQ adjusted), suspect IDs = 30.3%, fillers = 12.1%

descriptions of the perpetrator. Verbatim witness comments were significantly more often recorded in DB-SEQ lineup reports (82.6% vs. 38.9%), a qualitatively different type of statement than that reported in the NB-SIM lineup. Specifically, the DB-SEQ lineup reports were more likely to include the eyewitness's phrasing (e.g., "that's her, she's got gold teeth"), whereas NB-SIM lineup reports contained more third-person accounts of the eyewitness decision (e.g., "the victim made a positive ID"). It should be acknowledged that the administrator's report form for sequential lineups specifically asked for "any words used by witness in making identification," whereas the simultaneous lineup protocol did not request this information.[6]

A stepwise logistic regression analysis was conducted in order to determine the strongest predictor of whether a lineup fell into the NB-SIM or DB-SEQ condition, using witness encoding variables as predictors—whether the offender was known to the witness, delay, type of crime (violent or not), presence or absence of weapon, cross or same-race, witness as bystander versus victim. The only significant predictor in this equation was the last variable, whether the witness was a victim or bystander, Wald's X^2 (1, $N = 69$) = 3.85, $p = .05$. A crime victim was more likely to view a DB-SEQ lineup, a bystander to view a NB-SIM lineup. A second stepwise logistic regression analysis was conducted with the same predictors plus lineup format, and the dependent measure of suspect identification. Not surprisingly, witnesses who knew the offender generated significantly more suspect picks, Wald's X^2 (1, $N = 73$) = 10.31, $p = .001$. With that variable removed from the list of predictors, same-race and bystander status become significant predictors of suspect identifications, Wald's X^2 (1, $N = 73$) = 4.75, $p = .03$, and Wald's X^2 (1, $N = 73$) = 4.07, $p = .04$, respectively. Same-race crime incidents resulted in 63% suspect identifications, cross-race incidents in 39% suspect identifications, Z_p ($N = 74$) = 2.03, $p = .04$. (Keep in mind that this dataset includes both known and stranger offenders.) Witnesses who were bystanders produced 72% suspect identifications, whereas victims produced 46% suspect identifications, Z_p ($N = 70$) = 1.86, $p = .06$. In the smaller subset of data for first identifications of strangers, bystander status remained a significant predictor of suspect identification, Wald's X^2 (1, $N = 43$) = 3.84, $p = .05$.

[6] The nine coding discrepancies stem from the following three cases. (1) The report for six lineups (from the same case) did not indicate lineup format, therefore the author excluded these from the current analysis. However, the independent coders noted that in the absence of a reported procedure, one could assume a non-blind simultaneous lineup and add these lineups to the analyses (5/6 involved non-stranger offenders and all resulted in suspect IDs). (2) In another case, a store clerk reported that she had seen the offender once a month earlier, but she subsequently demonstrated confusion about this point. This identification attempt was categorized by the author as a stranger–perpetrator, but independent coders suggested that the two (non-blind simultaneous) lineups should be categorized as verification lineups. If the eight lineups described above are reassigned as suggested, there would be an increase in NB-SIM lineups (6), to 42 in the "Evanston" column of Table 2, with revised witness response rates of 64.3% suspect IDs, 14.3% "ruled out," 14.3% "no choice," and 7.1% unreported. The "Adjusted" column of Table 2 would change minimally, now with 17 lineups and eyewitness response rates of 52.9% suspect IDs, 17.6% "ruled out" and 29.4% "no choice." More importantly, the addition of seven verification lineups accumulated from these two cases increases the percentage of NB-SIM verification lineups (Table 1) to 21 of 42 (50%), significantly more than the DB-SEQ lineup condition, Z_p ($N = 88$) = 2.80, $p = .005$. Differences between NB-SIM and DB-SEQ lineups (Table 1) also increase to statistically significant values for same-race identifications ($p = .02$) and for witness race ($p = .02$). (3) In one more case, the file is unclear as to whether a photo show-up involved the suspect who

Footnote 6 continued
appeared in the subsequent sequential lineup. If so, this would add one confirmatory lineup to the DB-SEQ tally. The difference between DB-SEQ and NB-SIM formats in number of confirmatory lineups would still be significant, $p = .03$.

Given the predictive value of same-race crimes and bystander status for suspect identification, additional analyses examined each of these two factors in the context of other witness encoding variables: delay, status of offender as known or stranger, presence/absence of a weapon, same versus cross-race, type of crime, and number of perpetrators. The lineups were first separated by same-race versus cross-race crime events. Same-race crimes were significantly more likely than cross-race crimes to involve a weapon, 63% vs. 33%, Z_p ($N = 74$) = 2.50, $p = .01$; multiple perpetrators, 43% vs. 17%, Z_p ($N = 71$) = 2.43, $p = .01$; and to produce a lineup within 24 hours of the crime, 38% vs. 17%, Z_p ($N = 74$) = 2.02, $p = .04$. Same-race crimes involved a significantly lesser proportion of Caucasian witnesses than did cross-race crimes, 13% vs. 80%, respectively, Z_p ($N = 76$) = 5.58, $p < .000$, and were somewhat (non-significantly) less likely to involve strangers, 63% vs. 75%, Z_p ($N = 66$) = 1.44, $p = .14$.

The lineups were next separated into bystander versus victim groups for comparison. No significant differences between bystanders and victims were found on the tested list of witness encoding variables. However, bystanders were somewhat less likely than victims to have been involved in a violent crime, Z_p ($N = 73$) = 1.43, $p = .15$, 61% vs. 78%, respectively.

Sequential laps. A pertinent question regarding field lineups is the impact of allowing the witness a second review "lap" of the sequential lineup (Klobuchar et al., 2006). In the set of 34 sequential lineups (Table 2), 20 of the witnesses requested a second lap. There was a small non-significant gain of two more suspect identifications in final decisions (from 24% to 29% first to final decision) and no change in the filler selection rate.

Analysis 2: Filler Quality

The evaluation of lineup fairness involves two related qualities: lineup bias and lineup size (Malpass & Lindsay, 1999). The suspect must not stand out on the basis of position in the lineup, photo quality, or other non-memory cue. In addition, members of the lineup must be reasonably plausible alternatives based on the description provided by the witness. Metrics of lineup fairness are offered in the literature and can be calculated for each of these lineups and then averaged across the sample set (see Brigham, Meissner, & Wasserman, 1999; Malpass, Tredoux, & McQuiston-Surrett, 2007; Tredoux, 1998; Wells & Bradfield, 1999 for useful discussion of the strengths and weaknesses of each index of lineup fairness).

Lineup bias is most simply calculated as the proportion of mock witnesses who choose the suspect (Doob & Kirshenbaum, 1973). One array among the 19 Evanston lineups was a noticeable outlier, 2.92 standard deviations above the mean, with 63.6% of the mock witnesses choosing the suspect. This lineup was a non-blind sequential presentation and was eliminated from subsequent calculations. From this point, ten double-blind sequential lineups produced a mean proportion of suspect identifications of 16.79, and seven non-blind simultaneous lineups produced a mean proportion of suspect identifications of 17.79. Neither of these figures differs significantly from the 16.67% expected by chance nor do they differ significantly from one another, Z_ps < 1. The combined 18 lineups (including one lineup with unreported lineup format) yield a 17.74 mean proportion of suspect picks ($SD = 10.86$, median = 16.70, range from 0.0 to 38.5). Conversely stated, an average 82.26% of mock witnesses chose a lineup member other than the suspect.

The number of members appearing in the lineup regardless of their physical appearance is referred to as *nominal size*. *Adjusted nominal size* can be calculated by subtracting from each lineup's nominal size the number of lineup alternatives that received no picks from mock witnesses (Malpass, 1981). Across the 18 lineups, the mean-adjusted nominal size was 5.67. In 78% of the lineups, all six photos drew at least one pick from the mock witnesses. These two figures—adjusted nominal size and proportion of mock witnesses choosing the suspect—provide basic evidence that the lineup fillers were at least plausible alternatives in the majority of the lineups, and that on average no significant bias against the suspect is evident in the lineup structure.

Additional indexes of lineup fairness can be calculated. If the computed *functional size* for each individual lineup is held to a maximum of six (a conservative value), the mean functional size of the 18 lineups is 5.05 (Wells & Bradfield, 1999). Mean *effective size* (Malpass, 1981) is 4.1, suggesting some truncation of lineup size. However, this reduction in lineup size did not result in high numbers of suspect picks in all cases; specifically, a filler was the favored pick for mock witnesses in 15 of the 18 lineups (83%). Assessment of *suspect bias* (Malpass, 1981) revealed a mean adjusted chance estimate of 24%, higher than the actual mean of suspect picks (17%). In summary, these calculations converge to offer evidence that the fillers were functioning well in this sample of 18 lineups, drawing mock witness responses in reasonable proportions. Although the lineup sizes may have been somewhat truncated, suspect bias was not elevated above chance level. The exception to this pattern was in the one non-blind sequential lineup discussed above.[7]

[7] Different viewpoints exist on what number of members constitutes a fair lineup. Brigham, Ready, and Spier (1990) consider a lineup fair if the effective size is equal to or greater than half the original nominal size of the lineup. Malpass (1981) argues for an effective size of at least five as the criterion for a fair 6-person lineup.

Discussion

The Evanston case files provide an *empirical* basis for comment on three aspects of the Illinois Pilot Program: the meaning of the overall results, the role of witness and suspect history in lineup outcomes, and the impact of a failure in random assignment of lineups to conditions. More broadly, the Evanston cases underscore the complexity of field lineup research and prompt recommendations for future field and lab investigation.

What Was Measured in the Illinois Program?

The Evanston lineup files offer a partial answer to this query. The quality of Evanston lineups was satisfactory and thereby unlikely to have inappropriately elevated suspect identification rates or lowered filler picks. However, the now-apparent verification and confirmatory lineups in the Evanston files (and presumably in Chicago and Joliet) place the Illinois eyewitness data in a new light. That is, higher suspect identification rates and lower filler selection rates can be expected compared to field or lab tests in which verification and confirmatory lineups are excluded. The contribution of these lineups to the Illinois outcomes is a relevant consideration as lineup practices between Illinois and other jurisdictions are compared and as the implications of the Mecklenburg Report are sorted for courtroom and policy decisions.

Witness and Suspect History

Repeated identifications are of concern beyond simple counting clarity. In Illinois, police may develop eyewitness evidence by securing identification of a suspect first through a photo lineup and subsequently with a live lineup, the same witness often providing both identifications. This procedure, which for some U.S. jurisdictions is favored practice, may appear reasonable at first glance and certainly necessary when a court generally precludes evidentiary use of photo lineups (e.g., *People v. Holiday*, 1970). Mecklenburg (2006b) claims that a live lineup occurs only after probable cause has been established and therefore that Illinois live lineups inherently have a greater rate of guilty suspects than do photo arrays. The implication is that live lineups have a higher suspect identification rate because the offender is more often in the lineup. At the same time, all Evanston confirmatory lineups were physical lineups.[8] The contribution of a previous photo identification to probable cause determination in these cases is unknown. Moreover, the second suspect identification by the same witness presents a substantial problem of interpretation when proffered as an indicator of eyewitness memory for the perpetrator of a crime, either for lineup field research or for trial evidence.

The witness's decision at a second identification task is fraught with serious confounds that challenge the fidelity of the witness's memory for the culprit. For example, the witness simply may detect that the suspect is the common denominator across two lineups. The fact that all fillers have changed from photo to live array is highly suggestive. An additional problem is that a witness who made a suspect pick at the first lineup may feel obligated to confirm the earlier pick at a second lineup even in the absence of true recognition memory (Dysart, Lindsay, Hammond, & Dupuis, 2001). Finally, people are much better able to recognize faces than they are able to remember where they saw them (Brown, Deffenbacher, & Sturgill, 1977). A memory failure for the circumstances of the previous encounter can fuel a witness's sense that a face in the second lineup is familiar although the correct context for that memory has been lost. Recognition of the suspect's face at a live lineup may stem from exposure to the suspect at the first array rather than at the crime scene. Evidence in support of these principles is found in laboratory work on a related topic, the *mugshot exposure effect* reviewed by Deffenbacher, Bornstein, and Penrod (2006). More recently, laboratory examination of the effects of repeated lineups (Steblay & Benson, 2009; Steblay & Tix, 2007) has found that risk to an innocent suspect increases sharply at the second lineup. This potentially damaging identification practice deserves more research attention.

Within-Condition Analyses

Beyond the aggregate numbers, the Illinois Program can be evaluated at two levels, within-condition and between-conditions. First, within-condition: Mecklenburg et al. (2008a) assert that the Program produced meaningful information about how lineups operate in the field. This point is correct in part; for example, the reported incidence of crime type in the Illinois jurisdictions (specifically sexual assaults) has been used as an indication of the limited availability of corroborating DNA evidence in the Illinois cases (Wells & Quinlivan, 2009). The Illinois Program fell short, though, in fully delineating police lineup practice. Now, the Evanston files offer more thorough information regarding characteristics associated with lineup operation, and the descriptive information of Tables 1 and 2 can be useful as jurisdictions and courts weigh the Evanston data and, by extension, the full Illinois Program. A few points should be kept in mind, however, if there is an attempt to move beyond description toward speculative explanation for lineup outcomes.

[8] However, not all physical lineups were confirmatory; 33% of physical lineups were confirmatory lineups.

First, the co-occurrence (multicollinearity) of factors within cases makes cause-and-effect relationships between case details and witness responses indecipherable. In Evanston, multicollinearity is evident, for example, in the comparison of same-race crimes to those in which the witness and perpetrator were of different races: same-race crimes were more likely to involve weapons and multiple perpetrators, and less likely to involve Caucasian witnesses, stranger–perpetrators, and delays between crime and lineup—a web of factors that are difficult to untangle. In the broader lineup research literature, authors of descriptive field studies have attempted to examine the impact of crime incident features on eyewitness decisions (Steblay, 2007b; Tollestrup, Turtle, & Yuille, 1994; Wright & McDaid, 1996). Yet these researchers prudently point out the dangers of pseudo-experimental comparisons. For example, weapon presence may be confounded with crime type (robbery versus fraud) and therefore also with factors such as differential witness attention, quality of culprit description, and delay prior to lineup. Although substantial support has been found in controlled experimental laboratory tests for the negative impact on memory of such factors as weapon presence, delay, and cross-race identification, descriptive field studies are likely to offer inconsistent and perhaps misleading results. Mecklenburg et al. (2008a) attempt to challenge laboratory findings regarding several variables—eyewitness age, weapon focus, and cross-race identification—but problems with these pseudo-experimental comparisons have become salient with the Evanston data and undermine claims of "far-reaching implications of this field data" (2008a, p. 26). The current FOIA dataset offers intriguing questions for future research, particularly as to the role of bystander versus victim status as a contributor to eyewitness accuracy, but cannot establish a cause–effect relationship between these two variables.

A second barrier to understanding the Evanston/Illinois data is the unknown impact on eyewitness decisions of non-blind administration and of relative judgment—the two potential intrusions on eyewitness memory that double-blind and sequential reforms, respectively, were developed to address (see, e.g., Clark, Marshall, & Rosenthal, 2009; Lindsay & Wells, 1985; Steblay et al., 2001). Higher suspect identification rates and fewer filler picks are expected from these two influences, particularly when circumstances combine to increase a witness's tendency to guess (Greathouse & Kovera, 2009). Furthermore, lack of blind administration in the simultaneous lineup condition may exacerbate the problem of repeated identification attempts; it can reasonably be expected that administrator influences that affect the first identification procedure (e.g., post-identification feedback; Douglass & Steblay, 2006; Wells & Bradfield, 1998) will carry over to a later identification attempt by the same witness. Thus, the impact of multiple lineup attempts may be uneven between double-blind sequential and non-blind simultaneous conditions.

Between-Condition Analyses

The Evanston data expose a serious failure of experimental control that is the undoing of any comparison *between* the two tested groups—lack of effective random assignment to the two lineup conditions of NB-SIM and double-blind sequential lineups. This point is critical. The Illinois study has been criticized on grounds of methodological theory, that is, on the *presumption* that non-random assignment to lineup format conditions made the playing field uneven to begin with (Wells, 2006b). The Evanston data now provide *direct evidence* that substantial and serious a priori differences existed between the two lineup conditions and that these differences led to circumstances that favored the non-blind simultaneous lineup over the double-blind sequential lineup.

In the face of the Schacter et al. (2008) judgment regarding the fundamental design confound of the Illinois study, Mecklenburg et al. (2008a) have tried to diminish concerns about non-blind administration of the simultaneous lineups by positing that non-blind administrator influence is nil. Now there is an additional reason for skepticism regarding the Illinois comparison of lineup conditions, one that is compelling because there is simply no means to reassure that this design error does not matter. When random assignment has been violated, the list of confounding variables that may differentially affect the two conditions, individually or in combination, cannot be exhaustively catalogued, let alone remedied with a variable-by-variable argument. The myriad ways in which eyewitness decision outcomes can be influenced—not only by eyewitness memory, but also by lineup procedures and investigation practices—can only be controlled by random assignment that evenly distributes the influence of these factors across tested conditions. Importantly, random assignment must be used to balance the occurrence of culprit-absent lineups between tested conditions.

The potential for problems of non-random assignment is also apparent in Chicago and Joliet. In each city, two sites were defined, by detective district and geographical region, respectively. Cases originating in one area were tested with non-blind simultaneous lineups and cases from another were tested with double-blind sequential lineups (Mecklenburg, 2006a, pp. 25–26). That is, although assignment to condition was predetermined and fixed (not at the discretion of the detective), the conditions themselves were defined on a non-random basis.

Investigative Practice and the Meaning of Field Outcomes

The Illinois Pilot Program prompts us to consider the meaning of field outcomes for an environment in which there is rarely ground truth. In the laboratory, the unit of analysis is an eyewitness's decision regarding a single lineup for a single crime at one point in time, and the correctness of that decision is immediately known; error is a function of the witness. In the field, the lineup is conducted within a different set of circumstances (Wells & Olson, 2002). Error can be a function of the witness or of an investigation that unknowingly has presented the witness with a culprit-absent lineup. This circumstance leads to two avenues for further consideration and research, and each is relevant to our understanding of field outcomes.

Scientists have long argued that blind lineup procedures are more likely to yield specific and fine-grained reports more faithful to true witness responses (Wells, 1988). One of the voiced suspicions about the very low filler pick rate (zero) in NB-SIM lineups in Evanston and Chicago was that the lineup administrators did not differentiate witness filler picks from no-choice responses (Wells, 2006a, 2008). The Evanston files do not provide direct evidence on this point. However, *only* the double-blind sequential lineup protocol *required* the detective to differentiate witness responses—a pick, no choice, or all ruled out—without knowledge of the suspect's identity or position in the lineup. This difference in lineup protocol allows the possibility that detective knowledge of the suspect influenced reporting of filler picks—and recording of lineup outcomes—differently between the two lineup conditions. Furthermore, the Evanston NB-SIM lineups show a significantly greater prevalence of detective paraphrasing of witness responses and of not recording the lineup outcome—the first field data consistent with the proposition of non-blind administrator recording effects. Of course, the design of the study prohibits a conclusion regarding a direct causal relationship between non-blind lineup administrations and recording lapses. More to the point, however, a blind administrator was simply not in a position to interpret the witness's pick or to slant the witness's comments toward an investigative hypothesis, because the blind administrator did not know the identity of the suspect.

A related factor that has received scant if any research attention is the impact on detective decision-making of a witness's qualifier for his/her decision. This qualifier may be in the form of a certainty statement ("I'm pretty sure, not positive") or a comment specific to a lineup member ("The hair was different, but I remember the shape of the face.") or both. The detective will memorialize the lineup results and determine the meaning of each witness's decision for the investigation and evidence record. Consider two lineups in which the witness has qualified his identification of the suspect. The first—"I'm 75% sure"—the detective records as tentative, a non-identification; the second—"He looks most like the person"—is recorded as a positive identification. How do we reconcile these recording decisions? More broadly, how do the type of qualifier and the detective's interpretation of the witness's statement affect the continuing investigation and evidence brought to trial? Conversely, do evidentiary standards dictate detective decisions (e.g., does any witness certainty statement of less than 100% unacceptably jeopardize the prosecution's case)? The Evanston data show that interpretation and recording of qualifiers differ under conditions of non-blind simultaneous versus double-blind sequential lineup delivery. The factor of non-blind administration is of particular importance in that the non-blind administrator not only knows who the suspect is, but may also be aware of the lineup decisions of other witnesses and of police possession of additional forms of evidence. This may feed into confirmation bias and damaging tunnel vision that will inhibit a self-correcting flow of criminal investigations (see Findley & Scott, 2006), including impacts on witness decisions if this information is shared (Hasel & Kassin, 2009). Future research should explore whether witness qualifiers are treated differently depending on aspects of police practice, such as lineup format, blind administration, photo versus live lineups, or stage of investigation.

The Evanston case files are especially instructive in revealing one additional avenue for future inquiry: analysis of lineups at the level of the investigation rather than the single lineup. Full case files alert researchers to the identification history of witness and perpetrator, as discussed above. Beyond this, an often lengthy crime investigation is revealed. Unlike the single point in time tested in the lab, police leads are refined or discarded and evidence is accumulated toward resolution of the question at hand: Who did it? Even the simplest assumption at the level of a single lineup (e.g., the criminal was a stranger) may be revealed to be erroneous as the investigation continues (the witness at first withheld information about his previous association with the perpetrator). Additional evidence, perhaps DNA, may definitively rule out a suspect, rendering a witness's previous identification "wrong" or a rejection of an early lineup "right." In cases that involve a serial criminal, a number of eyewitnesses attempt to identify various suspects during an investigation that over time comes to recognize that multiple cases are connected and then zeroes in on a single suspect. Review of the investigation chronology (and hindsight) may reveal the correctness of witnesses' early "failures" to identify the suspect from culprit-absent lineups. In short, aggregate outcomes from individual lineups blur the assessment of ground truth because investigations err—at times the suspect is not the perpetrator. Finding ground truth becomes a task of determining the correctness not only of the eyewitness's decision regarding a *charged* suspect (perhaps by weighing the strength of accumulated evidence), but also the correctness of the investigation's hypothesis regarding *each* suspect-as-perpetrator.

Forensically relevant recommendations for future field and laboratory research stem from this analysis of Evanston lineups: (1) that researchers establish ground truth in part through close tracking of field investigations across time; (2) that laboratory research explores the impact on eyewitness accuracy of repeated lineups for the same witness and suspect; (3) that laboratory investigations elucidate the meaning of finer-grained witness responses and qualifiers, and determine their links to identification accuracy; and (4) that both field and lab examine the impact of lineup practice, witness responses, and evidentiary standards on detectives' evaluation, recording, and pursuit of eyewitness evidence.

Summary

To return to the phrasing of Cutler and Kovera (2008), we now have a better understanding of what we are arguing about. The FOIA lawsuit has brought relevant information, even absent cooperation from Chicago and Joliet. Interpretation of field results demands clarity about underlying street practice and study design. Analysis of the Evanston files has improved our understanding of both, with new facts that bear on old questions about lineup quality and about each witness's exposure to the offender, identification task experience and response to the lineup. Eyewitness scientists now have field evidence in support of methodological theory. The Evanston files provide empirical evidence that the strategy used to assign lineups to the two tested conditions was not effective and that this non-random assignment resulted in a set of a priori circumstances that favored the non-blind simultaneous condition. Real-world data now also document the manner in which witness responses are recorded in non-blind simultaneous lineups versus double-blind sequential lineups, recording differences consistent with the proposition that non-blind simultaneous lineups have undesirable consequences for the documentation of witness responses. Law enforcement now has increased information for stringent comparisons of Illinois procedures to local practice. And, importantly, policy-makers can and should recalibrate their conclusions about the Illinois Program with a more complete recognition of the methodological flaws that prevent the Illinois Program from providing valid comparisons between lineup formats or from offering grounded challenges to laboratory findings. Finally, analysis of the FOIA files underscores the absolute necessity that field lineup experiments employ double-blind procedures and true random assignment.

Acknowledgment The author thanks Kassi Bloom for her work in the evaluation of Evanston lineup quality.

References

Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior, 25*, 475–491.

Brigham, J. C., Meissner, C. A., & Wasserman, A. W. (1999). Applied issues in the construction and expert assessment of photo lineups. *Applied Cognitive Psychology, 13*, S73–S92.

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

Brown, E., Deffenbacher, K., & Sturgill, W. (1977). Memory for faces and the circumstances of encounter. *Journal of Applied Psychology, 62*(3), 311–318.

Capital Punishment Reform Study Committee Act, 2003, Ill. Ill. Comp.Stat. Ann. 725 5/107A-10 (West 2006).

Clark, S. E., Marshall, T. E., & Rosenthal, R. (2009). Lineup administrator influences on eyewitness identification decisions. *Journal of Experimental Psychology: Applied, 15*, 63–75.

Cutler, B. L., & Kovera, M. B. (2008). Introduction to commentaries on the Illinois Pilot Study of Lineup Reform. *Law and Human Behavior, 32*, 1–2.

Deffenbacher, K. A., Bornstein, B. H., & Penrod, S. D. (2006). Mugshot exposure effects: Retroactive interference, mugshot commitment, source confusion, and unconscious transference. *Law and Human Behavior, 30*, 287–307.

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups—Partial remembering. *Journal of Police Science and Adminstration, 18*, 287–293.

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.

Dysart, J. E., Lindsay, R. C. L., Hammond, R., & Dupuis, P. (2001). Mug shot exposure prior to lineup identification: Interference, transference, and commitment effects. *Journal of Applied Psychology, 86*(6), 1280–1284.

Findley, K. A., & Scott, M. S. (2006). The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review, 2*, 291–397.

Greathouse, S. M., & Kovera, M. B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior, 33*, 70–82.

Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications? *Psychological Science, 21*, 122–126.

Kings County Supreme Court. (2008). Defendant's request to modify double-blind simultaneous lineup to sequential one denied. Matter of Derrick Walthour. *New York Law Journal, 239*, 30, col. 3, March 14.

Klobuchar, A., Steblay, N., & Caligiuri, H. L. (2006). Symposium: reforming eyewitness identification: Convicting the guilty, protecting the innocent: Improving eyewitness identifications: Hennepin County's blind sequential pilot project. *Cardozo Public Law, Policy, & Ethics Journal, 4*, 381–413.

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential presentation. *Journal of Applied Psychology, 70*, 556–561.

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

Malpass, R. S., & Lindsay, R. C. L. (1999). Measuring lineup fairness. *Applied Cognitive Psychology, 13*, 1–7.

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup fairness. In R. C. L. Lindsay, D. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology (Vol. 2): Memory for people* (pp. 155–178). Mahwah, NJ: Lawrence Erlbaum & Associates.

Malpass, R. S., Zimmerman, L. A., Ross, S. J., Topp, L. D, Uribe, V., De Leon, D., … Belisle, J. (2006a, March). Illinois Lineup Evaluation Project Protocol and Code Book. Appendix 16 in Mecklenburg, S. H. (2006, March). *Report to the legislature of the State of Illinois: The Illinois Pilot Program on sequential double-blind identification procedures*. Springfield, IL: Illinois State Police.

Malpass, R. S., Zimmerman, L. A., Ross, S. J., Topp, L. D., Uribe, V., De Leon, D., … Belisle, J. (2006b, March). Results of the Illinois Lineup Project. Appendix 17 in Mecklenburg, S. H. (2006, March). *Report to the legislature of the State of Illinois: The Illinois Pilot Program on sequential double-blind identification procedures*. Springfield, IL: Illinois State Police.

Mecklenburg, S. H. (2006a). *Report to the legislature of the State of Illinois: The Illinois pilot program on double-blind, sequential lineup procedures*. Springfield, IL: Illinois State Police.

Mecklenburg, S. H. (2006b). *Addendum to the report to the legislature of the state of Illinois: The Illinois Pilot Program on sequential double-blind identification procedures*. June 19, 2006, p. 9.

Mecklenburg, S. H., Bailey, P. J., & Larson, M. R. (2008a). The Illinois field study: A significant contribution to understanding real world eyewitness identification issues. *Law and Human Behavior, 32*, 22–27.

Mecklenburg, S. H., Bailey, P. J., & Larson, M. R. (2008b). Eyewitness identification: What chiefs need to know now. *The Police Chief, 75*(10), 68–81.

National Association of Criminal Defense Lawyers. (2007). Complaint for declaratory and injunctive relief. In the Circuit Court of Cook County Chancery Department, County Division, National Association of Criminal Defense Lawyers, plaintiff, v. Superintendent of the Chicago Police Department, Chief of the Evanston Police Department, Chief of the Joliet Police Department, and Director of the Illinois State Police NO. 07 CH 3622.

People v. Holiday, 47 Ill.2d 300, 307, 265 N.E2d 634, 637 (1970).

Ross, S. J., & Malpass, R. S. (2008). Moving forward: Response to "Studying eyewitness investigations in the field". *Law and Human Behavior, 32*, 16–21.

Schacter, D. L., Dawes, R., Jacoby, L. L., Kahneman, D., Lempert, R., Roediger, H. L., et al. (2008). Policy forum: Studying eyewitness investigations in the field. *Law and Human Behavior, 32*, 3–5.

Steblay, N. (2006). Observations on the Illinois Lineup Data. Retrieved from http://www.augsburg.edu/psychology/steblay.html.

Steblay, N. (2007a). Affidavit of Nancy Steblay. In the Circuit Court of Cook County Chancery Department, County Division, National Association of Criminal Defense Lawyers, plaintiff, v. Superintendent of the Chicago Police Department, Chief of the Evanston Police Department, and Director of the Illinois State Police, Defendants, No. 07 CH 03622, September 24.

Steblay, N. K. (2007b). Double-blind sequential police lineup procedures: Toward an integrated laboratory and field practice perspective. Final report to the National Institute of Justice, Grant #2004-IJ CX-0044. Retrieved from http://www.augsburg.edu/psychology/steblay.html.

Steblay, N. K. (2008). Commentary on "Studying eyewitness investigations in the field": A look forward. *Law and Human Behavior, 32*, 11–15.

Steblay, N., & Benson, S. (2009). *Multiple identification attempts with simultaneous and sequential lineups*. Paper presented at American Psychology-Law Society Conference. San Antonia, TX.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25*, 459–473.

Steblay, N. K., & Tix, R. (2007). *Multiple identification attempts with simultaneous lineups*. Paper presented at John Jay College of Criminal Justice Conference, Off the witness stand: Using psychology in the practice of justice. New York City.

Tollestrup, P. A., Turtle, J. W., & Yuille, J. C. (1994). Actual victims and witnesses to robbery and fraud: An archival analysis. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments*. New York: Cambridge University Press.

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

Wells, G. L. (1988). *Eyewitness identification: A system handbook*. Toronto: Carswell Legal Publications.

Wells, G. L. (2006a). Eyewitness identification: Systemic reforms. *Wisconsin Law Review*, *2*, 615–643.

Wells, G. L. (2006b). Gary L. Wells comments on the Mecklenburg report. Retrieved from http://www.psychology.iastate.edu/faculty/gwells/Illinois_Project_Wellscomments.pdf.

Wells, G. L. (2008). Field experiments on eyewitness identification: Towards a better understanding of pitfalls and prospects. *Law and Human Behavior, 32*, 6–10.

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*(3), 360–376.

Wells, G. L., & Bradfield, A. L. (1999). Measuring the goodness of lineups: Parameter estimation, question effects, and limits to the mock witness procedure. *Applied Cognitive Psychology, 13*, 27–39.

Wells, G. L., Luus, C. A. E., & Windschitl, P. (1994). Maximizing the utility of eyewitness identification evidence. *Current Directions in Psychological Science, 3*(6), 194–197.

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7*(2), 45–75.

Wells, G. L., & Olson, E. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied, 8*, 155–167.

Wells, G. L., & Quinlivan, D. S. (2009). Suggestive eyewitness identification procedures and the Supreme Court's reliability test in light of eyewitness science: 30 Years later. *Law and Human Behavior, 33*, 1–24.

Wright, D. B., & McDaid, A. T. (1996). Comparing system and estimator variables using data from real lineups. *Applied Cognitive Psychology, 10*, 75–84.