

Mock Jurors' Reactions to Selective Presentation of Evidence from Multiple-Opportunity Searches

Jonathan J. Koehler · William C. Thompson

Published online: 20 June 2006

© American Psychology-Law Society/Division 41 of the American Psychological Association 2006

Abstract Prior to trial, litigants sometimes conduct broad investigations in which there are multiple opportunities to find supportive evidence by chance alone. During trial, litigants may selectively present only the most helpful evidence uncovered by their investigations. Two experiments examined whether mock jurors appreciate that the evidence they hear at trial may be a selective and unrepresentative sample of underlying facts. The data suggest that people do understand the significance of multiple-opportunity searches for legal inference. However, they may not consider the possibility that evidence was strategically selected from a larger sample space of facts unless that sample space is identified.

Keywords Selection · Juror bias · Juror decision making · Biased evidence

Introduction

In criminal and civil cases, investigators frequently generate evidence by combing through large bodies of information for seemingly relevant facts. For example, investigators may screen records from banks, credit card agencies, telephone carriers, computers, or e-mail accounts in an effort to link a suspect to a particular location or event. Forensic scientists may examine numerous hairs, fibers, print impressions and biological samples from a crime scene looking for anything that could link a suspect to the crime. As technology continues to advance, and as we continue to seek new ways to increase our security, evidence generated through broad screening searches will become increasingly common.

J. J. Koehler (✉)

Behavioral Decision Making Faculty, McCombs School of Business,
The University of Texas at Austin,
1 University Station B6500, Austin, Texas
e-mail: koehler@mail.utexas.edu

W. C. Thompson

Department of Criminology, Law & Society, University of California,
Irvine, California, 92697-7080

One concern about broad searches is that they have the potential to generate seemingly relevant evidence by chance. Statisticians have long been aware of a related phenomenon that Diaconis and Mosteller (1989) playfully refer to as “The Law of Truly Large Numbers.” According to this Law, even extremely improbable events—such as “cancer clusters” (Fienberg & Kaye, 1991) and double lottery winners (Associated Press, 2002; Kolata, 1990)—are likely to occur if the sample space is very large. When one considers that millions of people have cancer, and millions of people buy multiple lottery tickets, the real surprise would be if there were no coincidental cancer clusters or double lottery winners (see e.g., Samuels & McCabe, 1986).

The significance of this point for civil and criminal cases is that broad investigations may turn up circumstantial evidence that *appears* to be highly probative of a hypothesis (e.g., “the suspect is guilty”) but which is actually coincidental. This problem seems particularly acute when a search is conducted in a manner that creates multiple opportunities to find support for a particular hypothesis. We refer to such searches as *multiple-opportunity searches*. Because multiple-opportunity searches create a greater chance of finding hypothesis-supportive facts by chance alone (i.e., even when the hypothesis is false), a fact may be less meaningful if it is discovered in a multiple-opportunity search than if the same fact is uncovered in a single-opportunity search. However, if the trier-of-fact is not told whether incriminating evidence is the product of a single- or multiple-opportunity search, he or she may be confused about how much weight that evidence deserves. An interesting question, and one that we address in two experiments later, is how people treat evidence that *may* be the product of a multiple- opportunity search relative to how they treat evidence that they know to be the product of either a single- or multiple-opportunity search.

Selection bias at trial

A key feature of the Anglo-American system of justice is that litigants control the production and presentation of evidence (Damaska, 1997). As a result, the evidence heard by juries is not a representative sample of information uncovered by the parties. Instead, it is a selective sampling of facts that is biased to suit the interests of the litigants. We refer to the strategic (i.e., nonrepresentative) sampling and presentation of information obtained from a multiple-opportunity search as a *selection bias*.¹

Some scholars, such as Froeb and Kobayashi (1996), acknowledge that the accuracy of legal verdicts could suffer if parties are permitted to mislead juries by selectively reporting favorable evidence. However, these authors also contend that selection bias is not a serious problem in court because the adversarial process allows parties to expose limitations and omissions in their opponent’s evidence, and provides strong economic incentives for them to do so. By this account, jurors ultimately hear appropriately balanced presentations, notwithstanding the tendency of each party to present evidence selectively. We disagree for several reasons.

First, parties do not always have equal investigative resources or equal access to critical information. Consequently, the selective presentations of one party may not be exposed and corrected. For example, attorney–client privilege and the privilege against self-incrimination limit the government’s access to information from the defense in criminal cases. Likewise, the defense commonly receives limited cooperation from government witnesses and has limited access to important investigative resources such as government databases and police files.

¹ Others have used the term “selection bias” in a similar way. Berk (1983) states that selection bias occurs when “potential observations from some population of interest are excluded from a sample on a nonrandom basis” (p. 390). Faigman, Kaye, Saks, and Sanders (2002) define selection bias as a sample that has been “drawn in a way that makes it unrepresentative of the population to which are inferences are to be made” (p. 129).

A second reason that jurors may ultimately hear unbalanced presentations is that parties may choose not to divulge how broadly they searched to obtain their evidence. Even in criminal cases, where the government has a constitutional obligation to disclose both incriminating and exculpatory evidence uncovered during the police investigation (Brady & Maryland, 1963), it is not clear whether the government must disclose information about the breadth of its search or about failed investigative leads. Yet such information may provide valuable clues about the probative value of the incriminating evidence that was found. In short, when parties selectively present facts culled from information to which they have superior access, the jury is likely to hear a selective and unbalanced presentation of two different sets of evidence, rather than a balanced presentation of the evidence as a whole.

Jurors undoubtedly realize that they are hearing competing one-sided presentations. Nevertheless, they may have trouble using this knowledge to draw accurate conclusions (Brenner, Koehler, & Tversky, 1996). Research shows that people have difficulty reasoning with probabilistic evidence when the sample space and sampling process that generate the evidence are unclear (Fiske & Taylor, 1991; Nisbett, Krantz, Jepson, & Kunda, 1983; Nisbett & Ross, 1980). To the extent that jurors remain ignorant about the pool of information from which evidence presented at trial is selected, they too may draw inappropriate conclusions. Research also indicates that people—including those with advanced quantitative training (Tversky & Kahneman, 1971)—draw overly strong conclusions from small samples of evidence (Kahneman & Tversky, 1972; Tversky & Kahneman, 1974). This judgment error occurs even when those evidence samples are known to be unrepresentative. Hamill et al. (1980) asked people to make judgments about American prison guards after seeing an interview with a single guard who was depicted as either humane or brutal. The interview had a powerful effect on people's judgments about prison guards in general, regardless of whether the guard was described as "typical" or "atypical" of other prison guards. Whether jurors' verdicts are similarly influenced by small, selected evidentiary samples is an open question.

To date, few studies have considered how people in the role of jurors react to evidence that has been—or may have been—selectively chosen from a multiple-opportunity search. In two experiments, we investigate whether people treat evidence that was generated by a broad, multiple-opportunity search differently than evidence generated by a narrow, focused search. We also explore how people treat evidence when they do not know the nature of the search that generated it.

Experiment 1

We examined this issue in a mock criminal trial experiment in which the police had identified a suspect before searching for additional evidence to link that suspect to the crime. The breadth of the search varied. In the *narrow search* condition, jurors were told that the police looked for a few items of evidence that might link the suspect to the crime and that they found each of those items. In the *broad search* condition jurors were told that the police looked for many possible items of evidence that might link the suspect to the crime and that they found evidence of a few specific links. In the *search of unknown breadth* condition (hereafter the "unknown condition"), jurors were told that the police found evidence of a few incriminating links, but the breadth of the police search that yielded this evidence was left unclear.

We predicted that jurors in the narrow search condition would give more weight to the case against the defendant than jurors in the broad search condition. Of greater practical and theoretical importance, perhaps, is how jurors in the unknown condition would respond. These jurors were in a position most analogous to that of actual jurors who receive evidence without

knowing much about the process that generated the evidence. We predicted that jurors in the unknown condition would ignore the possibility of selection bias. Operationally, this means that we expected that jurors in the unknown condition would respond to the evidence in the same way as would jurors in the narrow search condition. We base this prediction on research that shows people overvalue unrepresentative samples (Nisbett & Ross, 1980) and reason poorly with evidence when the sample space and sampling process are unclear (Fiske & Taylor, 1991; Nisbett et al., 1983).

Method

Participants

Three hundred and one jury-eligible University undergraduates were mock jurors (“jurors”) in this experiment. Jury-eligibility was inferred from responses to questions about citizenship, prior convictions, and age. Jurors received partial course credit in an introductory business class in exchange for their participation.

Design

This experiment employed a 3 (breadth of search: narrow, broad, or unknown) \times 2 (number of actual links between defendant and crime: 2 or 6) fully crossed between-participants design. We also included a *control condition* that did not receive any evidence about investigated leads or links between the suspect and the crime. Individual jurors judged the strength of the case and the probability that the defendant was guilty, and rendered a verdict (without deliberation).

Materials and procedure

Jurors read a written description of a case in Austin, Texas, in which a man allegedly broke into a woman’s apartment at night in the late 1980s and attempted to rape her. The woman fought with her assailant and scratched him before he fled the scene when neighbors responded to her screams. Police were unable to develop any suspects, and the case grew cold. Fifteen years later, the police used a new technology to develop a DNA profile of a blood sample that had been collected from beneath the woman’s fingernails on the night of the attempted rape. This profile was examined against a state-wide DNA database of criminal offenders and a single match was found with a man from the Dallas area. The frequency of the matching DNA profile was 1 in 10 million. However, the victim was unable to identify the suspect from a line-up. The police arrested the suspect and investigated further in hopes of uncovering additional evidence that linked him to the Austin crime.

When constructing stimulus materials, we drew upon a population set of 30 possible links between the suspect and the crime. We sampled links from this set to determine the investigated leads and actual links in each of the experimental conditions. The investigated leads fell into six categories: (a) clothing worn by the assailant (e.g., lizard skin cowboy boots), (b) vehicles observed leaving the scene (e.g., dark colored pickup truck), (c) characteristics of the assailant (e.g., mustache), (d) items the assailant displayed during the attack (e.g., handcuffs), (e) locations or events at which the assailant may have seen or met the victim (e.g., a beauty pageant), and (f) the defendant’s criminal history (e.g., assault with a knife). There were five possible

investigated leads within each category. To expand the generality of the study, and to minimize the risk that our results would depend on features of particular investigated and actual links, we created a unique stimulus set for each participant. Consequently, there was no group deliberation in this study.

We assigned jurors at random to one of six experimental groups or to a control group. Jurors in three of the experimental groups were told that police found *two* specific items of evidence that linked the suspect to the crime. Jurors in the other three experimental groups were told the police found *six* specific items of evidence that linked the suspect to the crime.

Within the two-item and six-item conditions we also varied what jurors were told about the number of investigated leads (i.e., the number of possible links between the suspect and the crime). In the *narrow search* conditions (which we refer to as “2-of-2” and “6-of-6”), the police investigated 2 (or 6) possible links and found evidence to support each of the possible links they investigated. Because the links were sampled from the stimulus population, each juror received a different set. For example, one juror in the 2-of-2 condition may have been told that the perpetrator wore a Chicago Bulls cap and carried a yellow plastic flashlight, and that the investigation revealed that the suspect had owned both items around the time of the crime, whereas another juror might have been told that the perpetrator wore a Led Zeppelin T-shirt and had a gold earring, and that the investigation revealed that the suspect had owned both of these items around the time of the crime.

In the *broad search* conditions (which we refer to as “2-of-30” and “6-of-30”), the police investigated 30 possible links between the suspect and the crime and found evidence to support 2 (or 6) actual links. Participants in these conditions were given a list of 30 leads that police investigated (these comprised the entire population of possible links) and were told that police had found evidence to confirm 2 (or 6) of these links. We used a random process to determine which possible links police confirmed such that each juror received a unique set. Jurors were given no further information about the leads that police were unable to confirm. In the *unknown* conditions (which we refer to as “2-of- x ” and “6-of- x ”), the police found evidence to support 2 (or 6) actual links after investigating an unknown number of possible links. We refer to the 2-of- x and 6-of- x conditions as the unknown conditions because jurors were not told and did not know whether the evidence they received was the product of a narrow or broad investigation. Hence, as far as they knew, the evidence might (or might not) have been selected from a broad, multiple-opportunity search.

Jurors in the *control* condition did not receive any information about leads the police investigated or actual links police found between the suspect and the crime. They were told only about the DNA match and the victim’s failure to identify the matching man in a line-up.

Results

Table 1 presents the results of Experiment 1. Overall, the results indicated that jurors discounted (i.e., attached less weight to) the case against the defendant when the evidence resulted from a broad search relative to when it resulted from a narrow search, but did not discount the case when the evidence resulted from a search of unknown breadth.

Strength of case

An ANOVA of jurors’ case strength judgments detected main effects for breadth of search, $F(2, 259) = 22.35, p < .001$, and number of links, $F(1, 259) = 21.76, p < .001$, as well as a significant interaction between breadth of search and number of links, $F(2, 259) = 5.15, p < .01$. The

interaction arose because the breadth of search variable produced stronger effects in the 2-link conditions than in the 6-link conditions. In the 2-link groups, planned contrasts showed that judgments of case strength in the broad search condition ($M = 5.42$) were significantly lower than those in either the narrow search condition ($M = 8.1, p < .01$; Cohen's $d = 1.20$) or the unknown search condition ($M = 7.7, p < .01$; Cohen's $d = 1.04$). The narrow and unknown conditions did not significantly differ. The same pattern appeared in the 6-link groups although the effect sizes were smaller. Planned contrasts showed that judgments in the broad search condition ($M = 7.5$) were significantly lower than those in either the narrow search condition ($M = 8.3; p < .05$; Cohen's $d = 0.41$) or the unknown search condition ($M = 8.5; p < .05$; Cohen's $d = 0.50$). Again, the narrow and unknown conditions did not significantly differ. Jurors in the 6-link groups may have been somewhat impressed with the discovery of such a large number of links between the crime and the defendant regardless of the breadth of the search that produced the incriminating evidence.

Probability of guilt

The pattern of results on probability of guilt estimates was nearly identical to the pattern on case strength judgments (see Table 1). An ANOVA detected main effects for breadth of search, $F(2, 261) = 21.82, p < .001$, and number of links, $F(1, 261) = 20.30, p < .001$, as well as a significant interaction between these two variables, $F(2, 261) = 4.22, p < .05$. As before, the interaction arose because the magnitude of the differences among breadth of search conditions was smaller in the 6-link groups than in the 2-link groups (see Table 1).

In the 2-link groups, planned contrasts showed that probability of guilt judgments in the broad search condition ($M = 56.9$) were significantly lower than those in either the narrow search condition ($M = 86.0, p < .01$; Cohen's $d = 1.18$) or the unknown search condition ($M = 78.3, p < .01$; Cohen's $d = 0.84$). The narrow and unknown conditions did not significantly differ. The same pattern appeared in the 6-link groups although, again, the effect sizes were smaller. Planned contrasts showed that judgments in the broad search condition ($M = 78.4$) were significantly lower than those in either the narrow search condition ($M = 88.1; p < .05$; Cohen's $d = 0.48$) or the unknown search condition ($M = 89.3; p < .05$; Cohen's $d = 0.53$). Again, the narrow and unknown conditions did not significantly differ.

Table 1 Mean judgments and conviction rates as a function of number of actual links and search condition (Experiment 1)

Number of actual links/nature of search	<i>n</i>	Case strength	Probability of guilt	Verdict (% guilty)
Two links				
Narrow: 2-of-2	30	8.1	86.0	83.3
Unknown: 2-of- <i>x</i>	43	7.7	78.3	74.4
Broad: 2-of-30	45	5.4	56.9	35.6
Six links				
Narrow: 6-of-6	43	8.3	88.1	74.4
Unknown: 6-of- <i>x</i>	49	8.5	89.3	88.0
Broad: 6-of-30	53	7.5	78.4	69.6
Control	38	7.1	79.1	42.1

Note. Case strength ratings range from 1 (not at all strong) to 10 (extremely strong). Probability of guilt estimates are percentages.

Verdict

Finally, a binomial logistic regression analysis on the verdict dependent measure produced results that were roughly similar to those described earlier. As before, there was a significant interaction between breadth of search and number of links, Wald $\chi^2(2) = 7.17$, $p < .05$. In the 2-link conditions, contrast analyses revealed a main effect for breadth of search in which jurors in the broad search condition ($P = 35.6\%$) were less likely to convict than jurors in either the narrow search condition ($P = 83.3\%$; $p < .001$) or the unknown condition ($P = 74.4\%$; $p < .001$), whereas conviction rates in the narrow and unknown conditions did not significantly differ.

In the 6-link conditions, most jurors (77.2%) voted to convict. The only significant difference that emerged using contrast analyses was that jurors in the broad search condition ($p = 69.6\%$) convicted less frequently than those in the unknown condition ($p = 88.0\%$; $p < .05$).

Control group

The judgments and verdicts of jurors in the control group provide a baseline from which to judge the impact of the investigation evidence on jurors. Recall that jurors in this group did not receive evidence against the defendant other than the DNA match and the victim's failure to identify him. Data that appear at the bottom of Table 1 indicate that jurors in the control group regarded the common evidence to be moderately strong ($M = 7.1$) and thought that there was a high probability that the defendant was guilty ($M = 79.1\%$). However, only 42.1% of jurors in this group returned a guilty verdict. Post hoc contrast analyses revealed that jurors in the control group assigned higher case strength ratings and probability of guilt judgments than jurors in the 2-of-30 condition ($p < .01$ and $p < .001$, respectively). This result indicates that production of additional evidence against a defendant can actually weaken a prosecution's case when that evidence is produced by a broad search. Post hoc analyses also indicated that the control group convicted less frequently than all experimental groups ($p < .01$ on all) except the 2-of-30 group. This result confirms previous reports that mock jurors are reluctant to return guilty verdicts when there is little or no direct evidence that confirms a reported DNA match (Koehsler, Chia, & Lindsey, 1995).

Discussion

The central purpose of this experiment is to investigate how people treat evidence that was (or may have been) generated by a broad, multiple-opportunity search. Jurors in the broad search conditions (where police had 30 opportunities to find links between the defendant and the crime) were less impressed by the case against the defendant than jurors in the narrow search conditions (where police found the same number of links but had fewer opportunities to find links in the first place).

One explanation for this finding is that jurors in the broad search condition gave weight to the investigated leads that police failed to confirm. Here, we note that the failure of police to find support for possible links does not logically imply that the existence of the link was disproved or even that the case against the suspect is significantly weaker than it would otherwise be. The absence of evidence that the suspect had owned a Chicago Bulls baseball cap 15 years earlier, for example, does not disprove the hypothesis that the suspect owned such a cap. Nevertheless, jurors in the obvious selection condition may have thought that the existence of so many failed

investigative leads was diagnostic of innocence. A second, related, explanation for the finding is that jurors may have been influenced by the ratio of investigated leads to confirmed links, and felt that a confirmation rate of 2 in 2 (or 6 in 6) was more diagnostic of guilt than a rate of 2 (or 6) in 30.

A third explanation is that people appreciate the statistical principle that underlies selection bias when the relevant sample space is called to their attention. According to this explanation, jurors in the broad search conditions were less impressed with the case against the defendant than jurors in the narrow search conditions because they understood that some links between the defendant and the crime might have been found by chance alone given the large number of leads that the police pursued. This explanation is also consistent with our findings concerning how jurors evaluated the case when they did not know whether the evidence was generated by a broad or narrow search. We found that these jurors treated the evidence exactly as did those who knew it came from a narrow, single-opportunity search. We explain this phenomenon by suggesting that jurors who were not informed about the search process lacked access to crucial sample space information and therefore could not access the statistical principle that was available to those who knew that the links resulted from a broad, multiple-opportunity search.

A multiple-opportunity search will usually go hand-in-hand with the existence of failed investigative leads. Broad searches produce many opportunities to find incriminating links as well as many opportunities to fail to find incriminating matches. Hence, regardless of which explanation best accounts for our data, it seems reasonable to conclude that jurors may be less impressed by a case when they know the evidence was generated by a broad search that turned up a few seemingly relevant facts than if they think the same evidence was uncovered in a narrower, targeted search.

The most interesting finding in Experiment 1 was that jurors who were ignorant about the breadth of the search viewed the case against the defendant in much the same way as those who knew that the evidence was produced from a narrow, targeted investigation. This finding has important implications for the way jurors treat evidence at trial. Although investigators often pursue many possible leads to link a target suspect to a crime, only those leads that produce evidence against the defendant are likely to be presented at trial. Indeed, judges may see little relevance in spending time revealing and discussing failed investigative leads. Without this information, however, jurors may give more weight to the case than they would have if they had a more complete understanding of the process by which the matching evidence was produced.

Experiment 1 used a large number of distinct stimulus sets. This was done to promote generalizability of findings across diverse evidentiary items. A cost of this approach is that we were not able to include group deliberation. Hence, a key question left unanswered by Experiment 1 is whether and how group deliberation affects this phenomenon. We examine this question in Experiment 2.

Experiment 2

Method

Participants

One hundred and five jury-eligible University undergraduates were mock jurors (“jurors”) in this experiment. Jurors participated in 18 groups of four to seven.

Materials and procedure

Jurors listened to an audiotape of a criminal trial and were provided a transcript of the audiotape so they could read along. In this case, the police identified a suspect for a 15-year-old rape through a “cold-hit” in a DNA database, and then searched for additional evidence that linked the suspect to the crime.

As in Experiment 1, jurors were randomly assigned to one of three selection conditions. In the narrow search (“2-of-2”) condition, police investigated two potential links between the suspect and crime and found evidence to support actual links in both instances. In the broad search (“2-of-30”) condition, the police investigated 30 possible links and found evidence of two actual links. As in Experiment 1, no further information was provided about investigated leads that failed to link the suspect to the crime. In the unknown (“2-of- x ”) condition, no information was provided about the extent of the search that generated evidence of two actual links. In all conditions, the suspect was arrested and went on trial.

Jurors individually judged the strength of the case against the defendant, estimated the probability of the defendant’s guilt, and rendered a preliminary verdict. Next, jurors deliberated on the case for up to 20 min or until the group reached agreement on a verdict. At that point they again provided individual judgments of case strength, probability of guilt, and a verdict. Deliberations were monitored through a one-way window and recorded on audiotape.

There were six replications of the three-cell design (i.e., six juries per condition). Each replication employed a separate stimulus set with a unique pair of links between the suspect and crime. We used a random process to select the pair of links in each replication. The links were selected from the same set of 30 links used in Experiment 1 with the constraint that no link was used more than once.

A pair of raters who were blind to the hypotheses and experimental conditions coded audiotapes of deliberations. The raters tallied each comment that related to the DNA evidence, the confirmed links between the defendant and the crime, or the number of leads that the police had investigated. Each comment was further classified as pro-prosecution, pro-defense, or neutral. Operationally, raters tallied comments into one of nine categories on a 3×3 matrix. To improve reliability of individual ratings, the raters discussed and agreed on the classification of each comment before it was tallied. A second team of raters jointly rescored comments made by six randomly selected juries. The ratings assigned by the two teams were generally consistent (Pearson $r = .90$).

Design

This experiment employed a fully crossed 3 (breadth of search: narrow, broad, or unknown) $\times 2$ (time of assessment: pre-deliberation, post-deliberation) mixed variables design. Search condition was a between-participants variable, and time of assessment was a within-participants variable. Jurors judged the strength of the case, estimated the probability that the defendant was guilty, and provided verdicts both before and after group deliberation.

Results

Table 2 presents the results of Experiment 2. As expected, jurors in the broad search condition gave lower ratings on strength-of-case and probability of guilt, and were less likely to vote guilty, than jurors in the narrow search and unknown search conditions. This effect was observed both before and after group deliberation.

Table 2 Mean judgments and conviction rates as a function of deliberation and search condition (Experiment 2)

Nature of search	<i>n</i>	Case strength	Probability of guilt	Verdict (% guilty)
Pre-deliberation				
Narrow: 2-of-2	33	7.4	77.0	63.6
Unknown: 2-of-x	37	7.2	75.3	67.6
Broad: 2-of-30	35	5.3	54.1	28.6
Post-deliberation				
Narrow: 2-of-2	33	8.1	86.2	81.8
Unknown: 2-of-x	37	8.4	89.2	86.5
Broad: 2-of-30	35	4.7	48.5	22.9

Note. Case strength ratings range from 1 (*not at all strong*) to 10 (*extremely strong*). Probability of guilt estimates are percentages.

Strength of case ratings were analyzed using a repeated measures ANOVA, which revealed main effects for search, $F(2, 102) = 25.92$, $p < .001$, and time of assessment (pre- or post-deliberation), $F(1, 102) = 5.64$, $p < .05$, and a Search \times Time of assessment interaction, $F(2, 102) = 8.87$, $p < .001$. The interaction arose because the difference between the broad search condition and the other two conditions became more pronounced after deliberation. On the pre-deliberation measures, planned contrasts showed that ratings in the broad search condition ($M = 5.3$) were significantly lower than the ratings in either the unknown condition ($M = 7.2$; $p < .01$; Cohen's $d = 0.87$) or the narrow search condition ($M = 7.4$; $p < .01$; Cohen's $d = 1.02$). The latter two conditions did not differ significantly. On the post-deliberation measures, the ratings in the broad search condition ($M = 4.7$) were again significantly lower than ratings in either the unknown condition ($M = 8.4$, $p < .01$, Cohen's $d = 1.69$) or the narrow search condition ($M = 8.1$, $p < .01$; Cohen's $d = 1.47$). Again, the latter two conditions did not differ.

Jurors' probability of guilt judgments paralleled their case strength ratings. A repeated measures ANOVA again revealed main effects for search condition, $F(2, 102) = 27.37$, $p < .001$, and for deliberation, $F(1, 102) = 8.47$, $p < .05$, as well as a significant interaction between search condition and deliberation, $F(2, 102) = 9.01$, $p < .001$. Again, this interaction arose because the gap between judgments in the broad search condition and the other two conditions was larger after deliberation than before deliberation. On the pre-deliberation measures, probability of guilt judgments in the broad search condition ($M = 54.1$) were significantly lower than ratings in either the unknown condition ($M = 75.3$; $p < .01$; Cohen's $d = 0.83$) or the narrow search condition ($M = 77.0$; $p < .01$; Cohen's $d = 0.98$). The latter two conditions did not differ significantly. On the post-deliberation measures, judgments in the broad search condition ($M = 48.5$) were again significantly lower than ratings in either the unknown condition ($M = 89.2$, $p < .01$, Cohen's $d = 1.73$) or the narrow search condition ($M = 86.2$, $p < .01$; Cohen's $d = 1.59$). The latter two conditions did not differ.

The pattern for verdicts was similar. There were significant differences among the three search conditions on pre-deliberation verdict preferences, $\chi^2(2) = 13.04$, $p < .001$. These differences arose because the number of guilty verdicts was significantly lower in the broad search condition (28.6%) than in the narrow search condition (63.6%); $\chi^2(1) = 8.42$, $p < .01$, or the unknown search condition (67.6%); $\chi^2(1) = 10.95$, $p < .001$. The latter two conditions did not differ significantly. There was a similar pattern on post-deliberation verdicts, $\chi^2(2) = 38.29$, $p < .001$. Again, the number of guilty verdicts was significantly lower in the broad search condition than in the narrow search condition (22.9% and 81.8%, respectively; $\chi^2(1) = 23.64$, $p < .001$) or the

unknown condition, 86.5%; $\chi^2(1) = 29.49, p < .001$. Logit loglinear analyses supported these conclusions. The best-fitting model for post-deliberation verdicts posited a lower conviction rate in the broad search condition than the unknown or narrow search conditions, which did not differ from each other, $\chi^2(2) = 0.98, p > .05$.

During deliberations, which averaged 12.2 min (range: 3–20 min), jurors made more comments about the DNA evidence ($M = 15.5$) than about the evidentiary links between the defendant and the crime ($M = 7.8$). However, there were no significant differences among conditions in the number of times either factor was mentioned, nor were there differences in the proportion of comments about either factor coded as pro-prosecution, pro-defense, or neutral.

Jurors made only four comments about the number of opportunities police had to find links between the defendant and crime. Three of these comments occurred in two juries in the 2-of-30 condition; one occurred in a jury in the 2-of-2 condition. All four comments were pro-defense. Most juries did not discuss the number of opportunities the police had to find incriminating links between the defendant and the crime.

Discussion

Experiment 2 replicated and extended the results of Experiment 1. Jurors gave more weight to evidence generated by a narrow search (2-of-2 condition) and evidence generated by a search of unknown breadth (2-of- x condition) than to evidence they knew was generated by a broad search (2-of-30 condition). We observed this pattern both before and after group deliberation. Even after discussing the evidence with other jurors, it apparently did not occur to jurors in the unknown (2-of- x) condition that the incriminating evidence may have arisen from a broad investigation in which the police investigated many potential links. Indeed, none of the jurors in the potential selection condition mentioned the nature of the police search during deliberations.

As in Experiment 1, jurors gave less weight to the links between the defendant and the crime when they knew this evidence was the product of a broad investigation. Jurors who were explicitly told that the police pursued many leads in addition to those that yielded evidence against the defendant (2-of-30 condition) were much less impressed by the evidence than were jurors who were told that each lead the police pursued ended up incriminating the defendant. After deliberation, this gap between the selection and no selection conditions expanded. This result suggests that additional reflection on the evidence may have increased some jurors' sensitivity to the number of opportunities police had to find incriminating links.

General discussion and conclusion

Litigants sometimes discover helpful evidence by culling large bodies of information. Although this approach can uncover information that would otherwise be missed, evidence identified in this manner is potentially misleading in two ways. First, a broad search may create multiple opportunities to discover facts that appear to support a particular hypothesis but which are actually coincidental. Second, a broad, multiple-opportunity search may allow a litigant to present a selective sample of evidence that is unrepresentative of the full range of facts uncovered by the investigation. Our studies with mock jurors suggest that people understand the significance of multiple-opportunity searches when this feature of the search process is called to their attention. Jurors in Experiments 1 and 2 gave less weight to the prosecutor's case when they were told that the reported links were a small subset of the investigated leads than when the reported links resulted from the only leads that the police had investigated.

These experiments do not distinguish between two closely related explanations for why jurors are more skeptical of cases built on broad searches. The large number of investigated leads may have helped jurors appreciate the selective nature of the reported links—which made them less impressive as evidence against the suspect. Alternatively, jurors may have treated the investigated leads that did not link the suspect to the crime as exculpatory evidence. Regardless of which underlying explanation is at work, these findings should allay concerns about the potential for jurors to be misled by selective evidence from multiple-opportunity searches—at least where the nature of the search is disclosed.

However, our studies also suggest that people fail to think much about how evidence is generated unless they are told about it. When they were not told about the nature of the search that produced the evidence, jurors in our experiments treated the evidence as if it had been produced by a narrow search that looked only for the evidence that was found. The absence of information about the nature or scope of the investigation did not cause them to discount the case against the defendant. Apparently, these jurors did not consider the possibility that the evidence they received might have been cherry-picked from a larger, multiple-opportunity search.

Our results are consistent with Fiske and Taylor's (Fienberg & Kaye, 1991) claim that people have trouble reasoning with incomplete evidence samples, in part, because "questions of sample adequacy rarely arise in a person's mind" (p. 352). Our results are also consistent with data that show people are not well equipped for drawing inferences from the absence of information (Einhorn & Hogarth, 1978; Ward & Jenkins, 1965). As Hearst (Hearst, 1991) explains, "[H]uman beings and other animals have trouble using the mere absence of something as a basis for efficient and appropriate processing of information. They notice and recall additions much more readily than deletions" (p. 432). In these experiments, the mere mention of evidence leads that did not pan out amounted to an "addition" that enabled jurors to convert difficult-to-use nonevent information into readily used and well-understood event information.

From a policy standpoint, our findings also highlight the importance of disclosing the nature and scope of the investigations that produce the evidence presented at trial. Such disclosure would seem to be particularly important in cases involving multiple-opportunity searches. Without disclosure, jurors are unlikely to consider the possibility that the available evidence was strategically selected from a broader set of leads, and may give the evidence more weight than they otherwise would.²

Will jurors be told about multiple-opportunity searches?

Whether jurors always *will* be told about the scope and nature of searches is an open question. Much depends on whether expert witnesses, lawyers and judges appreciate the importance of disclosure. Although we hope this paper contributes to greater awareness of this issue, we worry that the same psychological tendencies that blind jurors to the implications of biased sampling may affect many experts, lawyers, and judges as well. A forensic scientist might testify, for example, that the defendant's hair matches a hair found at the crime scene without realizing the importance of explaining that the matching hair was only 1 of the 20 tested. Of course, opposing attorneys can and should raise questions during cross-examination about the search process that produced evidence against their clients. Even if they appreciate the importance of doing so,

² Sometimes the probative value of the evidence is so large that breadth of search makes little difference. Many DNA matches fall into this category.

however, they may face relevancy objections that require them to convince time-conscious and skeptical judges that the inquiry is worthwhile.

Whether parties are obligated, under current discovery rules, to disclose multiple-opportunity searches and selective presentation of evidence is an interesting legal question. It is not clear whether the current rules of discovery incorporate such a requirement. Although the government has a constitutional obligation to disclose exculpatory evidence uncovered by its criminal investigation (*Brady v. Maryland*, 1963), failed investigative leads will not necessarily be viewed as exculpatory because the failure to find evidence supporting a hypothesis does not necessarily disprove the hypothesis. A famous aphorism, well known among lawyers, holds that “absence of evidence is not evidence of absence.”

Even if failed investigative leads are seen as exculpatory, failure to disclose that information is not necessarily grounds for reversal in cases that include convictions. In (*Strickler v. Greene*, 1999), Justice Souter wrote that the materiality of withheld evidence turned on whether there was a “significant possibility” of a different outcome had it been disclosed to the defense (p. 298). Four years earlier, the Supreme Court described the standard as one of “reasonable probability” (*Kyles v. Whitley*, 1995, p. 433). With these standards in mind, a question arises as to whether the obligation to disclose extends to information about the scope of the investigation, such as the number or nature of failed investigative leads.

Some indication appears in the 10th Circuit’s ruling in the Oklahoma City bombing case against Terry Nichols (*U.S. v. Nichols*, 2000). Nichols argued that he was denied due process when prosecutors failed to disclose all 40,000 FBI “lead sheets” that were used to record information from potential sources of evidence in the case. At trial, the government disclosed only those lead sheets that produced information they deemed relevant. The 10th Circuit ruled against the Nichols on grounds that the verdict probably would not have been different if the facts contained in the undisclosed lead sheets had been presented to the jury. Although the court did not rule broadly on the disclosure obligation for failed leads, this ruling hints that the Brady requirement may not extend to all aspects of the evidence search process.

Without access to *evidence* on the scope of a party’s investigation, opposing attorneys might still cross-examine witnesses about the breadth of the search, and raise the possibility of cherry-picked evidence in arguments to the jury. Whether such tactics will increase jurors’ skepticism about potentially selected evidence, in the absence of specific information about the investigation, is another interesting topic for future research.

Acknowledgment

The authors thank Michael Kromer for his assistance with Experiment 2.

References

- Associated Press (2002). A win–win for man in Lotto, Fantasy 5. *Los Angeles Times*.
- Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, *48*, 386–398.
- Brady v. Maryland* (1963). 373 U.S. 83, 83 S. Ct. 1194, 10 L. Ed. 2d 215.
- Brenner, L. A., Koehler, D. J., & Tversky, A. (1996). On the evaluation of one-sided evidence. *Journal of Behavioral Decision Making*, *9*, 59–70.
- Damaska, M. (1997). *Evidence law adrift*. New Haven, CT: Yale University Press.
- Diaconis, P., & Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association*, *84*, 853–861.
- Inhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: persistence of the illusion of validity. *Psychological Review*, *85*, 395–416.

- Faigman, D. L., Kaye, D. H., Saks, M. J., & Sanders, J. (2002). *Science in the law: standards, statistics, and research issues*. St. Paul, MN: West Group.
- Fienberg, S. E., & Kaye, D. H. (1991). Legal and statistical aspects of some mysterious clusters. *Journal of the Royal Statistical Society A*, *154*, 61–74.
- Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed). New York: McGraw-Hill.
- Froeb, L. M., & Kobayashi, B. H. (1996). Naïve, biased, yet Bayesian: can juries interpret selectively produced evidence? *Journal of Law, Economics and Organization*, *12*, 257–276.
- Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: generalizing from atypical cases. *Journal of Personality and Social Psychology*, *39*, 578–589.
- Hearst, E. (1991). Psychology and nothing. *American Scientist*, *79*, 432–443.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.
- Koehler, J. J., Chia, A., & Lindsey, J. S. (1995). The random match probability (RMP) in DNA evidence: irrelevant and prejudicial? *Jurimetrics Journal*, *35*, 201–219.
- Kolata, G. (1990). 1-in-Trillion coincidence, you say? Not really, experts find. *New York Times*, p. c1.
- Kyles v. Whitley* (1995). 514 U.S. 419.
- Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*, 339–363.
- Nisbett, R. E., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.
- Samuels, S. M., & McCabe, G. P. (1986, February 27). More lottery repeaters are on the way. *New York Times*, p. A22.
- Strickler v. Greene* (1999). 527 U.S. 263, 298 (Souter, J. concurring).
- Tversky, A., & Kahneman, D. (1971). The belief in the 'law of small numbers'. *Psychological Bulletin*, *76*, 105–110.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*, 1124–1131.
- U.S. v. Nichols* (10th Cir. 2000). 242 F. 3d 391.
- Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology*, *19*, 231–241.