

Statistical Inference and Forensic Evidence: Evaluating a Bullet Lead Match

42

43

Suzanne O. Kaasa · Tiamoyo Peterson · Erin K. Morris ·
William C. Thompson

44

45

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

46

Abstract This experiment tested the ability of undergraduate mock jurors ($N = 295$) to draw appropriate conclusions from statistical data on the diagnostic value of forensic evidence. Jurors read a summary of a homicide trial in which the key evidence was a bullet lead “match” that was either highly diagnostic, non-diagnostic, or of unknown diagnostic value. There was also a control condition in which the forensic “match” was not presented. The results indicate that jurors as a group used the statistics appropriately to distinguish diagnostic from non-diagnostic forensic evidence, giving considerable weight to the former and little or no weight to the latter. However, this effect was attributable to responses of a subset of jurors who expressed confidence in their ability to use statistical data. Jurors who lacked confidence in their statistical ability failed to distinguish highly diagnostic from non-diagnostic forensic evidence; they gave no weight to the forensic evidence regardless of its diagnostic value. Confident jurors also gave more weight to evidence of unknown diagnostic value. Theoretical and legal implications are discussed.

47

48

49

50

51

52

53

54

55

56

57

58

59

Keywords

Q1

60

Forensic science is playing an increasingly important role in criminal trials. In order to link a defendant to a crime, prosecutors have presented expert testimony on a variety of forensic techniques such as DNA analysis, fingerprint comparisons, toolmark comparisons, and bullet lead analysis (Thompson & Cole, 2006; Faigman, Kaye, Saks, & Sanders, 2002). When presenting forensic evidence, experts typically testify that a characteristic associated with one sample (e.g., a sample from a crime scene) “matches” a characteristic of another sample (e.g., one taken from

61

62

63

64

65

66

S. O. Kaasa · T. Peterson · E. K. Morris
Department of Psychology and Social Behavior, University of California, Irvine, California, USA

W. C. Thompson (✉)
Department of Criminology, Law & Society, University of California, Irvine, California 92697, USA
e-mail: William.Thompson@uci.edu

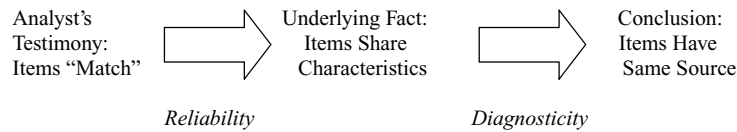


Fig. 1 Reliability and diagnosticity as distinct elements of the probative value of a forensic “match” for proving the matching items have a common source

67 the defendant), thereby suggesting that the samples may have a common source (Thompson &
68 Cole, 2006).

69 When considering the inferential tasks faced by jurors who must evaluate a forensic “match,”
70 it is helpful to distinguish two elements that David Schum and his colleagues have labeled
71 *reliability* and *diagnosticity* (Schum, 1994; Schum & DuCharme, 1971). As illustrated by Fig. 1,
72 the *reliability* of the forensic evidence is its value for proving an underlying fact: that two items
73 share a characteristic or set of characteristics. The *diagnosticity* of forensic evidence is the value
74 of the underlying fact (the shared characteristics) for establishing that two items have a common
75 source.

76 For example, jurors who hear that a defendant has the same DNA profile as a bloodstain found
77 at the crime scene must consider the *reliability* of this evidence. Do the DNA profiles really
78 match? Could there have been a laboratory mistake? Jurors must also consider the *diagnosticity*
79 of this evidence. Could this “match” be a coincidence? How likely is it that another person
80 would have the same DNA profile? When making these judgments, jurors sometimes consider
81 statistics (Thompson & Cole, 2006). For example, jurors may hear that the false positive rate of
82 a forensic test is less than 1% or that only one person in a thousand in a reference population has
83 the matching characteristic (Thompson, 1989). Ultimately, jurors must integrate probabilistic
84 judgments about both the *reliability* and *diagnosticity* of forensic evidence in order to assign it
85 weight (Thompson, Taroni, & Aitkin, 2003).

86 In order to learn more about how (and how well) people make such judgments, we examined
87 mock jurors’ evaluations of a particular type of forensic evidence: compositional analysis of bullet
88 lead (National Research Council, 2004). While we could have addressed our research questions
89 in connection with a variety of types of forensic evidence, bullet lead evidence was particularly
90 appropriate because normative models exist that specify how jurors *should* evaluate the strength
91 of bullet lead evidence (Thompson, 2005), and these models can be used as benchmarks for
92 human performance. Bullet lead evidence also has the advantage (for research purposes) of
93 being relatively obscure, which makes it unlikely that mock jurors’ reactions to it will be
94 influenced by preconceptions about its *reliability* or *diagnosticity* (as might be the case with
95 DNA evidence or fingerprints).

96 Bullet lead evidence

97 The FBI developed compositional analysis of bullet lead in the 1960’s and used the technique for
98 nearly 40 years to link bullets recovered from crime scenes to boxes of ammunition owned by
99 suspects (Finkelstein & Levin, 2005; Imwinkelried & Tobin, 2003). The lead alloy used to make
100 bullets contains trace levels of various elements, such as copper, arsenic, antimony, bismuth,
101 silver, cadmium and tin. FBI analysts used sophisticated instruments to measure the levels of
102 these trace elements in order to develop a chemical profile of each bullet. The FBI assumed
103 that bullets in a particular box are more likely to have originated in the same molten vat of lead

alloy and consequently more likely to have matching profiles than bullets from different boxes. Therefore, they testified that a bullet lead “match” was evidence that the bullets in question came from the same box (National Research Council, 2004). Bullet lead testing was particularly important in cases where the crime scene bullet was too damaged or too fragmented to compare to a particular gun, or where no gun was recovered.

Recently, a series of scholarly articles called into question the validity and probative value of bullet lead evidence (Randich, Duerfeldt, McLendon, & Tobin, 2002; Tobin & Duerfeldt, 2002). In a few criminal cases the admissibility of bullet lead evidence was successfully challenged (e.g., *United States v. Mikos*, 2003). As the controversy grew, the FBI commissioned the National Research Council (NRC) to convene a panel to study the scientific legitimacy of bullet lead evidence. In its report (National Research Council, 2004), the NRC panel called for some improvements in the FBI’s methods for “matching” the chemical profiles of bullets, but found the method overall to be *reliable*. The panel suggested, by way of illustration, that the hit rate of the FBI procedure—i.e., the probability of declaring a match between two samples from the same molten vat of lead—was approximately .90, while the false positive rate of the procedure—i.e., the probability of declaring a match between bullet from different vats—was only .002.

On the other hand, the NRC panel raised concerns about the *diagnosticity* of bullet lead evidence. It pointed out that a single vat of molten lead might be large enough to produce as many as 35 million “matching” bullets, and that these bullets might be distributed together through the supply chain such that large numbers of “matching” bullets could end up in a particular locality. While it might be possible to do research on the frequency of “matching” bullets in a particular area, relatively few studies of that type have been done. To complicate matters further, bullets from different molten vats (with different chemical profiles) sometimes are mixed during manufacturing and packaging, so that the bullets in a box of ammunition do not always match each other. The FBI’s research has shown that a single box of ammunition can contain bullets from as many as 14 distinct compositional groups (National Research Council, 2004, p. 5). Hence, when weighing bullet lead evidence, it is also important to consider what percentage of the defendant’s bullets “match” the bullet from the crime scene.

The NRC panel concluded that there is an insufficient scientific foundation at present to conclude that finding a “match” between bullets renders them likely to be from the same box. It strongly condemned expert testimony that suggests or implies that bullet lead evidence can link matching bullets to the same box. The panel recommended that, until further research is done, analysts instead limit themselves to saying that a bullet lead match renders it more likely that the matching bullets came from the same “compositionally indistinguishable volume of lead (CIVL).” Citing continuing concerns about the diagnostic value of bullet lead evidence, the FBI recently announced that it would discontinue bullet lead testing for the time being (Piller, 2005; Thompson, 2005). Because the FBI operated the only laboratory in the United States that did bullet lead testing, this decision effectively ended the use of the technique in American courts. Questions remain, however, about whether bullet lead evidence might have been misleading to juries in past cases (Pace, 2005; Piller, 2005; Imwinkelried & Tobin, 2003).

The normative question: How *should* jurors evaluate a forensic “match”?

Thompson (2005) presented a normative model of how jurors *should* evaluate bullet lead evidence. It provides a useful benchmark for assessing human performance. In this model, *SB* represents the hypothesis that the crime scene bullet and defendant’s bullet came from the *same box*; *DB* is the alternative hypothesis that the two bullets came from *different boxes*; *SC* is the event that the two bullets came from the *same compositionally indistinguishable volume of*

150 *lead (CIVL)*; *DC* that they came from *different CIVLs*; and *M* is the event that the laboratory,
 151 after performing bullet lead analysis, declares the bullets to *match*. Making certain simplifying
 152 assumptions, Thompson showed that the likelihood ratio describing the value of a bullet lead
 153 match for proving the matching bullets came from the same box is:

$$\frac{p(M|SB)}{p(M|DB)} = \frac{p(M|SC)p(SC|SB) + p(M|DC)p(DC|SB)}{p(M|SC)p(SC|DB) + p(M|DC)p(DC|DB)} \quad (1)$$

154 This model shows that the value of bullet lead evidence depends on four key statistical factors.
 155 The first two factors, which together determine the *reliability* of the evidence, are the hit rate,
 156 $p(M|SC)$, and false positive rate, $p(M|DC)$, of the analytical procedure. As noted earlier, the
 157 NRC report suggested that the hit rate might be .90 and the false positive rate .002. The third and
 158 fourth factors, which together determine the *diagnosticity* of the evidence, are the prevalence
 159 of matching bullets in the suspect's box, $p(SC|SB)$, and the prevalence of matching bullets in
 160 other boxes from which the crime scene bullet might have come, $p(SC|DB)$. For example, if
 161 half the bullets in the suspect's box match the crime scene bullet, $p(SC|SB)$ would be .50; if
 162 10% of the bullets in different boxes (i.e., possible source boxes other than the defendant's box)
 163 match the crime scene bullet, then $p(SC|DB)$ would be 0.10.¹

164 The empirical question: How *do* people evaluate a forensic match?

165 The goal of the present study was to explore whether people's intuitive assessments of bullet
 166 lead evidence correspond with the normative model. Because the main controversy over bullet
 167 lead evidence centers on its *diagnosticity*, we were particularly interested in whether people are
 168 sensitive to the statistical variables that affect *diagnosticity*. Do people recognize the difference
 169 between bullet lead evidence that is highly diagnostic and that which has little or no diagnostic
 170 value? Additionally, given the limited data available on the critical variables that affect *diagnos-*
 171 *ticity*, we wondered what people will do when asked to judge the value of bullet lead evidence
 172 in the absence of statistics on the key *diagnosticity* factors.

173 A number of previous studies have examined mock jurors' evaluations of statistical evidence.
 174 One important line of research has examined how mock jurors' evaluations of a forensic match are
 175 affected by statistics on the probability of a *false match* (Faigman & Baglioni, 1988; Goodman,
 176 1992; Smith, Penrod, Otto, & Park, 1996; Thompson & Schumann, 1987). Most studies have
 177 examined statistics on the probability of a coincidental match between individuals (or items)
 178 that happen by chance to share the matching characteristics. These statistics are often called
 179 random match probabilities (RMPs). Some studies also include statistics on the probability of
 180 a false match due to other factors, such as laboratory error (Koehler, Chia, & Lindsey, 1995;
 181 Nance & Morris, 2002; Schklar & Diamond, 1999) and evidence tampering (Nance & Morris,
 182 2005). These studies suggest that jurors generally respond to evidence of a forensic match by
 183 adjusting their judgments in an appropriate direction—the weight they give to a forensic match
 184 increases as the probability of a *false match* decreases. There is some evidence that people's
 185 judgments are “conservative”—i.e., that they give too little weight to evidence of a forensic

¹ The two remaining terms in Eq. (1) are simply complements of terms already defined. The term $p(DC|SB)$ refers to the probability the two bullets would arise from different CIVLs if they are from the same box. Within a given box the bullets will either *all* be from the *same CIVL (SC)* or they will be from one or more different CIVLs (*DC*). Because *DC* and *SC* are mutually exclusive and exhaustive events, $p(DC|SB)$ is simply $1 - p(SC|SB)$. By the same logic, $p(DC|DB)$ is the complement of $p(SC|DB)$.

match relative to Bayesian norms. However there is also evidence that people sometimes rely on fallacious forms of reasoning (Thompson, 1989; Thompson & Schumann, 1987) or simplifying heuristic strategies (Koehler & Macchi, 2004) that could cause them to over-value a forensic match.

The way in which statistical data are presented may also be important. Some researchers have concluded that people's evaluations of false match probabilities are, on the whole, reasonable and appropriate regardless of presentation format (Nance & Morris, 2005), while others have found that natural frequencies are better understood than probabilities and lead to better decisions (Thompson & Schumann, 1987; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000).

Although these studies are clearly relevant, they focused on an inferential task that may be easier than the task jurors face when evaluating some types of forensic evidence. In previous studies, the key variable that jurors needed to consider when evaluating the probative value of a forensic match was the probability of a *false match*—i.e., the probability of a “match” occurring if the matching items had a *different* source. Jurors did not need to consider the probability of a *true match*—i.e., the probability of a match occurring if the items had the *same* source—because a match was virtually certain if the items had the same source. For example, two blood samples from the same person are almost certain to have the same DNA profile. Hence, in these studies there was a simple and direct relationship between the probability of a *false match* and the probative value of the evidence: as the probability of a *false match* decreased, the probative value of the forensic “match” (for proving the matching items have the same source) always increased.

When jurors evaluate bullet lead evidence the situation is more complicated because the probability of a *true match* and a *false match* are both important variables. The probability of a *true match*, which is reflected in the numerator of equation 1, depends on both the hit rate of the test, $p(M|SC)$, and the probability that two bullets will have the same metal composition if they come from the same box, $p(SC|SB)$. Because bullets from different compositional groups often end up in the same box, the probability of a *true match* (i.e., the probability of a match occurring if the bullets are from the same box) may be *much* less than one. For example, it might be the case that only one bullet in ten in the defendant's box has the same composition as the crime scene bullet, suggesting $p(SC|SB)$ could be .10. If matching bullets were relatively common in the surrounding community, the probability of a *true match* might be as low, or even lower, than the probability of a *false match* (Thompson, 2005). Hence, it is possible that bullet lead evidence could have little or no probative value for proving two bullets came from the same box even if the probability of a false match is low (Thompson, 2005).

The study reported here examined mock jurors' sensitivity to the probative value of a forensic match when the probability of *true match* could vary. Jurors evaluated a hypothetical case in which the key evidence linking the defendant to a murder was testimony by a forensic expert that the murder bullet “matched” the metal composition of a bullet found in an open box of unexpended ammunition owned by the defendant. Jurors were assigned to one of four conditions. In the Strong Condition, statistical data were presented indicating that the “match” was highly diagnostic—i.e., $p(SC|SB)$ was high and $p(SC|DB)$ was very low. In the Worthless Condition, the statistical data indicated that the match had no diagnostic value—i.e., $p(SC|SB)$ and $p(SC|DB)$ were equally low. In the Unknown Condition, no statistical data were presented on $p(SC|SB)$ or $p(SC|DB)$. Finally, there was the Control Condition, in which the evidence of the bullet lead match was not presented. The goal of the study was to determine whether mock jurors distinguish evidence of a forensic match that is highly diagnostic from that which is non-diagnostic and to learn how they evaluate such evidence in the absence of statistical data about its diagnostic value.

235 Method

236 Procedure

237 Undergraduates ($N = 295$) recruited from a university human subjects pool participated in groups
238 of four to six. They received a written description of the evidence in a hypothetical criminal trial
239 and were asked to read along while listening to a 14-min recording of a narrator reading the
240 same description. When the narration was complete, participants (hereafter “jurors”) individually
241 responded to a brief pre-deliberation questionnaire. They were then left alone to deliberate on
242 the case for up to 20 min while an experimenter monitored the discussion from an adjoining
243 room. Then they individually responded to a post-deliberation questionnaire after which they
244 were debriefed, thanked and dismissed. The entire procedure took approximately 1 hr.

245 Materials and design

246 In the case described in the experimental materials, a robber entered a convenience store in a
247 small rural community, shot and killed the clerk with a .38 caliber handgun, and took money
248 from a cash register. Police later arrested the defendant because he matched an eyewitness’
249 description of the killer and because he had purchased .38 caliber ammunition from a local
250 Wal-Mart. Police recovered an open box of .38 caliber ammunition from the defendant, but no
251 gun (the defendant claimed his unloaded .38 caliber revolver had been stolen). The defendant
252 was unemployed, short of money, and owned a truck matching the description of one driven by
253 the killer.

254 In the Control Condition, jurors received no further evidence. In the three experimental
255 conditions, jurors were also told that an FBI examiner had conducted a bullet lead comparison and
256 had found a “match” between the murder bullet and a bullet from the defendant’s box. The jurors
257 in the experimental conditions received an extensive summary of the FBI examiner’s testimony,
258 which included a description of the bullet manufacturing process, the metal composition of
259 bullets, and the analytical tests that the FBI used to analyze the metal content of bullets. The
260 examiner’s testimony, which largely followed the recommendations of the 2004 NRC report,
261 included statistical data on the *reliability* of the FBI’s test—specifically, the *hit rate* of the FBI
262 test (“the probability that two bullets would match on the FBI test if they came from the same
263 CIVL is 0.90 (90%).”) and the *false positive rate* of the FBI’s test (“the probability of a match
264 by coincidence or error of two bullets from different CIVLs—the false positive probability—is
265 about 1 in 500 or 0.002.”).

266 The three experimental conditions varied with respect to the statistical data provided about
267 *diagnosticity*. In the Strong Condition, the FBI examiner testified that he had tested a random
268 sample of 20 bullets taken from the defendant’s box and that all 20 matched the murder bullet.
269 He also tested a “community sample” of 100 .38 caliber bullets collected shortly after the time
270 of the crime at a target range operated by a local gun club and found that none of these bullets
271 matched the murder bullet. By contrast, in the Worthless Condition, the FBI examiner testified
272 that only two of 20 bullets randomly sampled from the defendant’s box (10%) matched the murder
273 bullet and that 10 of 100 bullets from the community sample (10%) matched the murder
274 bullet. Finally, in the Unknown Condition, the FBI examiner gave no testimony about sampling
275 bullets from the defendant’s box or from the local community and gave no testimony about the
276 percentage of the defendant’s bullets or the percentage of bullets in the local community that
277 matched the murder bullet.

278 In all experimental conditions jurors heard the type of caveats about the *diagnosticity* of bullet
279 lead evidence that would be likely to emerge from effective cross-examination. Jurors were told

that the number of bullets produced from a single compositionally indistinguishable volume of lead (CIVL) might range from 12,000 to several million. They were also told that the National Research Council had been unable to obtain information about the probability of finding a large number of analytically indistinguishable bullets in a given geographic area, and that regional patterns in the distribution of ammunition are unclear.

Measures

The pre-deliberation questionnaire asked jurors for their “initial reactions” to the case “based on what you think right now.” The questionnaire asked them to rate the strength of the case against the defendant (on a 10-point scale anchored at “Not at all Strong” and “Extremely Strong”). It also asked them to estimate the “numerical probability” that the defendant was guilty by entering a number between 0 and 100%. Finally, it asked them to choose which verdict (guilty or not guilty) they thought they would return in the case if they were judging it as jurors and were instructed to convict only if the evidence convinced them beyond a reasonable doubt that the defendant was guilty.

The post-deliberation questionnaire began with the same three questions as the pre-deliberation questionnaire. As a manipulation check, jurors in the three experimental conditions were also asked whether they had received statistical data about the percentage of bullets in the defendant’s box and the percentage of bullets in a community sample that matched the murder bullet. They were also asked to make their own “best guess” estimate of each of those statistics.

Lastly, jurors completed a demographics questionnaire that included questions about their age, gender, prior jury service, criminal victimization, view of the death penalty, training in mathematics and statistics, and confidence in their ability to draw correct conclusions from numerical data.

Results

Table 1 presents the means of the main dependent measures in each experimental condition. We used STATA dummy-variable linear regression to compare results among conditions on the continuous measures (strength of case and probability of guilt) and STATA logistic regression to compare conviction rates. Because deliberation created intraclass correlations among juries, we analyzed post-deliberation results using clustered regression, clustering by jury.

Table 1 Pre- and post-deliberation judgments of case strength, probability of guilt and verdict by condition

Evidence ratings	Experimental condition			
	Strong ($n = 70$)	Worthless ($n = 73$)	Unknown ($n = 70$)	Control ($n = 82$)
Pre-deliberation				
Strength of case	6.99	6.58	6.49	5.84
Probability of guilt	70.29	63.47	62.71	57.07
Guilty verdict (%)	57	32	41	35
Post-deliberation				
Strength of case	6.81	5.47	5.79	4.79
Probability of guilt	70.03	56.13	56.81	49.55
Guilty verdict (%)	41	14	29	12

Note. Strength of case ratings range from 1 (not at all strong) to 10 (extremely strong). Probability of guilt is given as 1–100%. Guilty verdicts are the percent of jurors who voted guilty in each condition.

309 Strength of case

310 Condition was a significant predictor of pre-deliberation strength of case ratings, $R_{\text{adj}}^2 = .04$;
 311 $F(3, 290) = 4.81, p < .01$. Ratings in all three experimental conditions were higher than in the
 312 Control Condition (Strong vs. Control, $p < .001$; Worthless vs. Control, $p < .05$; Unknown vs.
 313 Control, $p < .05$). No other differences were significant.

314 Condition was an even better predictor of clustered post-deliberation ratings, $R_{\text{adj}}^2 = .12$;
 315 $F(3, 58) = 6.05, p = .001$, although the pattern of results across conditions was a bit different.
 316 Ratings in the Strong Condition continued to be higher than the Control Condition ($p < .001$) and
 317 ratings in the Unknown Condition were marginally higher ($p = .07$), but ratings in the Worthless
 318 Condition no longer differed significantly from the Control. Ratings in the Strong Condition
 319 were also higher than those in the Worthless Condition ($p < .01$) and marginally higher than
 320 those in the Unknown Condition ($p = .07$). No other differences were significant.

321 A change-score representing the difference between pre- and post-deliberation ratings was
 322 created for each juror. These scores, clustered by jury, were significantly predicted by condition,
 323 $R_{\text{adj}}^2 = .04; F(3, 58) = 2.93, p < .05$. Deliberation led to smaller changes in the Strong Condition
 324 than in the Worthless Condition ($p < .05$) or Control Condition ($p < .05$). No other differences
 325 were significant.

326 Probability of guilt

327 Condition was also a significant predictor of pre-deliberation probability of guilt estimates,
 328 $R_{\text{adj}}^2 = .04; F(3, 290) = 5.42, p < .001$. Estimates were higher in the Strong Condition than
 329 in the Worthless ($p < .05$), Unknown ($p < .05$) and Control Conditions ($p < .001$). Estimates
 330 in the Worthless Condition were also higher ($p < .05$), and those in the Unknown Condition
 331 were marginally higher ($p = .08$), than those in the Control Condition. No other pre-deliberation
 332 differences were significant.

333 For post-deliberation estimates, condition was again a significant predictor, $R_{\text{adj}}^2 = .11$,
 334 $F(3, 58) = 5.54, p < .01$. Estimates in the Strong Condition were significantly higher than rat-
 335 ings in the Worthless ($p < .05$), Unknown ($p < .05$) or Control Conditions ($p < .001$), but the
 336 Worthless and Unknown Conditions were no longer significantly different from the Control
 337 Condition.

338 Change-scores reflecting the difference between pre- and post-deliberation estimates varied
 339 by condition, although this relationship was only marginal, $R^2 = .03, F(3, 58) = 2.32, p = .08$.
 340 As with strength of case ratings, the changes following deliberation were smaller in the Strong
 341 Condition than in the Worthless Condition ($p < .05$) or Control Condition ($p < .05$). No other
 342 differences were significant.

343 Verdicts

344 Condition significantly predicted whether jurors voted guilty or not guilty before deliberation,
 345 $R_{\text{psuedo}}^2 = .03, \chi^2(3, N = 294) = 10.99, p = .01$. The conviction rate in the Strong Condition was
 346 higher than the Worthless Condition ($p < .01$), marginally higher than the Unknown Condition
 347 ($p = .06$), and higher than the Control Conditions ($p < .01$). No other differences were significant.

348 Results were similar for post-deliberation verdicts, $R_{\text{psuedo}}^2 = .07, \chi^2(3, N = 294) = 8.69$,
 349 $p < .05$. The conviction rate was again higher in the Strong Condition than in the Worthless
 350 Condition ($p < .05$) or Control Condition ($p < .05$), but no other differences were significant.

351 McNemar tests indicated that the conviction rate decreased significantly following delibera-
 352 tion in all conditions (all values significant at $p < .05$). However, logistic regression, clustering

by jury, found that the extent of the decrease in conviction rate was not significantly associated with experimental condition.

Recall of key statistics

As a manipulation check, we asked jurors following deliberation whether the FBI expert had presented statistics on the percentage of matching bullets in the defendant's box (*defendant match percentage*) and the percentage of matching bullets in a sample from the community (*community match percentage*). Among jurors in the Strong Condition and Worthless Condition, where the expert had presented those statistics, 75% correctly reported that they had received the *defendant match percentage* (16% incorrectly said they had not; the remainder said they did not know) and 72% correctly said they had received the *community match percentage* (about 21% incorrectly said they had not; the remainder said they did not know). In the Unknown Condition, where the expert did not present these critical statistics, only 37% correctly said they had not received the *defendant match percentage* (46% of jurors incorrectly reported they had received it, and 17% said "don't know") and 64% correctly said they had not received the community match percentage (29% incorrectly said that they had received it and 7% did not know).

Regression analyses were run in order to test whether jurors who answered both of these questions correctly gave a different pattern of responses on the main dependent measures than those who answered at least one of the questions incorrectly or said they didn't know. The results showed that this "correct recall" variable did not significantly predict any of the dependent measures (post-deliberation strength of case, $b = -.44$; probability of guilt, $b = -2.64$; guilt verdict, $b = -.25$, all *ns*).

Jurors' "best guess" estimates of the *defendant match percentage* varied significantly across conditions, $R^2 = .24$, $F(2, 42) = 27.94$, $p < .001$. The median estimate was 90% in the Strong Condition, 80% in the Unknown Condition and 30% in the Worthless Condition. Each condition differed significantly from the others (all p 's $< .01$), which indicates that our statistical manipulation successfully altered jurors' perceptions of this variable. However, some of the jurors may have been confused about what they were reporting. A surprising percentage (43% in the Strong Condition, 14% in the Worthless Condition, and 30% in the Unknown Condition) "guessed" the defendant match percentage was 90%. We suspect some of them were mistakenly reporting the "hit rate" of the FBI procedure, which was always .90.

Jurors' "best guess" estimate of the *community match percentage* also varied significantly across conditions, $R^2 = .15$, $F(2, 42) = 13.06$, $p < .001$, with each condition differing significantly from the others (all p 's $< .001$). The median estimates in the Strong Condition (4.5%) and Worthless Condition (10%) were close to the statistics the FBI agent provided, which provides further evidence that jurors perceived and were influenced by our manipulation of the key statistics. The median estimate in the Unknown Condition, where no community match percentage was provided, was much higher (50%) and may reflect the strong caveats jurors heard about the possibility (recognized by the NRC) of many matching bullets being found in the same community.

Implicit likelihood ratios

For each juror we computed an implicit likelihood ratio (ILR) by dividing their best guess estimate of the *defendant match percentage* by their best guess estimate of the *community match percentage* (for jurors who estimated the community match percentage to be zero we raised the estimate to 1 percent to avoid irrational numbers). The ILR provided an index of the extent to

397 which jurors thought a match was more likely if the bullet came from defendant's box (a *true*
398 *match*) than from another local source (a *false match*).

399 The ILR was a significant predictor of all pre- and post-deliberation responses on the
400 main dependent measures. A higher ILR was significantly associated with stronger ratings
401 of the strength of case against the defendant (pre-deliberation $R^2 = .02$, $b = .01$, $p < .05$; post-
402 deliberation, $R^2 = .05$, $b = .01$, $p < .01$), greater estimates of the probability of the defendant's
403 guilt (pre-deliberation $R^2 = .03$, $b = .09$, $p < .01$; post-deliberation $R^2 = .06$, $b = .13$, $p < .01$),
404 and a greater likelihood of finding the defendant guilty (pre-deliberation $R^2_{\text{pseudo}} = .04$, $b = .01$,
405 $p = .001$; post-deliberation $R^2_{\text{pseudo}} = .03$, $b = .01$, $p = .01$).

406 Interestingly, the numerator of the ILR (defendant match percentage) was a better predic-
407 tor than the ILR itself of strength of case (pre-deliberation $R^2 = .05$, $b = .01$, $p = .001$; post-
408 deliberation $R^2 = .12$, $b = .02$, $p < .001$), probability of guilt (pre-deliberation $R^2 = .10$, $b = .19$,
409 $p < .001$; post-deliberation $R^2 = .12$, $b = .23$, $p < .001$), and guilt verdicts (pre-deliberation
410 $R^2_{\text{pseudo}} = .10$, $b = .02$, $p < .001$; post-deliberation $R^2_{\text{pseudo}} = .08$, $b = .02$, $p < .001$). The de-
411 nominator (community match percentage) did not significantly predict responses for any of
412 these measures.

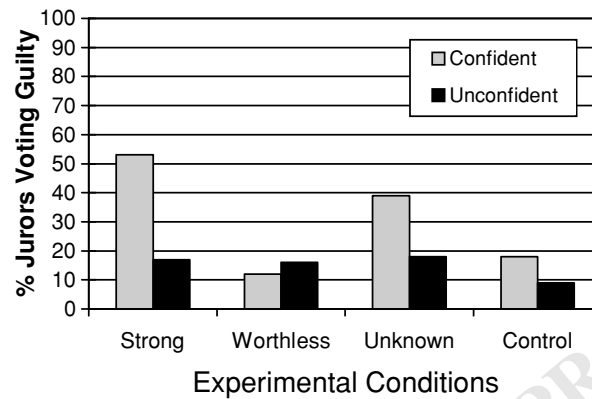
413 Individual differences

414 We used regression analyses to explore whether jurors' responses to the demographic ques-
415 tions were associated with their post-deliberation responses to the main dependent measures
416 (strength of case, probability of guilt and verdict). Although the regression models were signif-
417 icant (strength of case, $R^2 = .07$, $p < .05$; probability of guilt, $R^2 = .07$, $p < .05$; guilt verdict,
418 $R^2_{\text{pseudo}} = .05$, $p < .05$), the only individual predictor that contributed significantly to the model
419 was the question that asked: "How confident are you in your ability to draw correct conclusions
420 from numerical data, such as probabilities and frequencies?" More confident jurors gave higher
421 estimates of strength of the case ($b = .41$, $p < .01$) and probability of guilt ($b = 4.56$, $p = .001$),
422 and were more likely to render a guilty verdict ($b = .41$, $p = .01$). Interestingly, although res-
423 sponses to the confidence question were positively correlated with a measure of quantitative
424 training—i.e., the amount of mathematical and statistical training jurors had received in college
425 coursework ($R^2 = .05$, $b = .13$, $p < .001$), this latter measure was not itself a significant predictor
426 of any of the main dependent measures. Nor was there a significant relationship between the
427 confidence question and the accuracy of jurors' responses to questions about the statistical data
428 that the FBI expert had presented.

429 To further explore the effects of *confidence* on jurors' judgments, we divided jurors into
430 two groups—Confident and Unconfident—based on a median split of responses on the con-
431 fidence question. Among Confident jurors condition was a significant predictor of verdicts
432 both pre-deliberation ($\chi^2(3, N = 148) = 20.72$, $p < .001$) and post-deliberation ($R^2_{\text{pseudo}} = .11$,
433 $\chi^2(3, N = 148) = 10.59$, $p < .05$). Jurors in the Strong Condition gave significantly more
434 guilty verdicts than those in the Control Condition ($p < .05$) and Worthless Condition
435 ($p < .01$). In addition, jurors in the Unknown Condition were significantly more likely to
436 give guilty verdicts than jurors in the Worthless Condition, $p < .05$. No other differences were
437 significant.

438 Among Unconfident jurors, the conviction rate was generally lower and did not vary signifi-
439 cantly across conditions either before deliberation ($R^2_{\text{pseudo}} = .004$, $\chi^2(3, N = 138) = .73$, ns) or
440 after deliberation ($R^2_{\text{pseudo}} = .01$, $\chi^2(3, N = 138) = 1.01$, ns). Figure 2, which presents the pattern
441 of post-deliberation verdicts for Confident and Unconfident jurors, shows that differences among
442 conditions are accounted for almost entirely by Confident jurors.

Fig. 2 Post-deliberation percentage of Confident and Unconfident jurors voting guilty by condition



We also performed separate regressions within each experimental condition (clustered by jury) using confidence as a predictor for verdicts. Confidence was a significant predictor of guilty verdicts in the Strong Condition ($R^2_{\text{pseudo}} = .10$, $b = 1.74$, $p < .001$) and a marginally significant predictor in the Unknown Condition ($R^2_{\text{pseudo}} = .04$, $b = 1.05$, $p = .06$). Confidence was not a significant predictor for either the Worthless Condition ($R^2_{\text{pseudo}} = .004$, $b = -.34$, ns) or Control Condition ($R^2_{\text{pseudo}} = .02$, $b = .77$, ns).

Discussion

The independent variable in this experiment was the diagnostic value of the bullet lead match. This variable was manipulated by changing the data that jurors received about the percentage of matching bullets in the defendant's box and in a community sample.² The manipulation checks showed that the majority of jurors in these conditions correctly remembered having received the relevant data and that their "best guess estimates" of the key variables corresponded generally to the data presented. Accordingly, we concluded that the experiment created a fair test of jurors' ability to draw reasonable conclusions from statistical data on the diagnostic value of the "match."

The group means shown in Table 1 suggest that jurors readily perceived the difference in probative value between the highly diagnostic match (Strong Condition) and the non-diagnostic match (Worthless Condition). On the three main dependent measures (strength of case, probability of guilt, and verdict) ratings were consistently higher both before and after deliberation in the Strong Condition than in the Worthless Condition or Control Condition. These findings suggest that jurors do understand and are sensitive to statistical data on the diagnostic value of a "match," although this conclusion must be qualified in light of an internal analysis (discussed more fully below) which showed that the effect is accounted for almost entirely by a subset of "confident" jurors.

² In terms of the normative model presented in Equation 1, jurors in the Strong Condition received data indicating that $p(SC|SB)$ was very high and $p(SC|DB)$ was very low, and hence that the "match" was highly diagnostic. Jurors in the Worthless Condition received data indicating that $p(SC|SB)$ and $p(SC|DB)$ were equal, and hence that the "match" was non-diagnostic.

467 While jurors readily perceived the strength of the highly diagnostic “match,” they may not
468 have fully appreciated the weakness of the non-diagnostic “match,” at least not at first. Before
469 deliberation, jurors in the Worthless Condition gave significantly higher ratings of strength of
470 case and probability of guilt than jurors in the Control Condition, which suggests they were
471 giving *some* weight to the non-diagnostic match, albeit much less weight than jurors in the
472 Strong Condition were giving to the highly diagnostic match. After deliberation, the differences
473 between the Worthless and Control Conditions was no longer significant, but that may reflect the
474 lower statistical power of the clustered analysis, as the differences in group means were about
475 the same. On the other hand, jurors in the Worthless Condition had about the same conviction
476 rate as those in the Control Condition both before and after deliberations. Hence, to the extent
477 jurors perceived value in the non-diagnostic match, that perception was not reflected in their
478 verdicts.

479 In many actual cases evidence of a forensic match is not accompanied by statistical data on
480 its diagnostic value (Thompson & Cole, 2006). Jurors in those cases presumably must rely on
481 intuition and common sense to judge the diagnostic value of the evidence. To explore how jurors
482 respond to a forensic “match” in the absence of data on its diagnostic value, this experiment
483 included the Unknown Condition, where the expert presented data about the reliability of the
484 forensic match (hit rate and false positive rate of the test), but presented no data on the percentage
485 of matching bullets in the defendant’s box or in the surrounding community. Judgments in this
486 group fell roughly between those in the Strong Condition and those in the Worthless and Control
487 Conditions, suggesting that in the absence of data on diagnostic value jurors gave *some* weight
488 to this evidence, although not as much weight as they gave to the strong (highly diagnostic)
489 evidence.

490 Although our jurors successfully distinguished highly diagnostic from non-diagnostic forensic
491 evidence, it is impossible to determine, in the present experiment, whether their judgments
492 followed from the *ratio* of the *defendant match percentage*, $p(SC|SB)$, and the *community match*
493 *percentage*, $p(SC|DB)$, as specified in Equation 1. A possible alternative explanation is that
494 they relied solely (or primarily) on the defendant match percentage without taking into account
495 the community match percentage. This alternative explanation was supported by the finding that
496 jurors’ estimates of strength of the case, probability of guilt, and verdicts were correlated with
497 their “best guess” estimates of the percentage of matching bullets in the defendant’s box, but not
498 with their estimates of the percentage of matching bullets in a community sample. The alternative
499 explanation is also consistent with previous research showing that, when asked to evaluate the
500 impact of a datum, D , on the likelihood of a particular hypothesis, H , people often express far
501 more interest in knowing the probability of the datum *if the hypothesis is true*, $p(D|H)$, than the
502 probability of the datum if the hypothesis is false, $p(D|\bar{H})$, a phenomenon that has been called
503 the “pseudo-diagnosticity” effect (Beyth-Marom & Fischhoff, 1983; Doherty, Mynatt, Tweney,
504 & Schiavo, 1979). In order to definitively test this alternative explanation, future experiments
505 could simultaneously vary the defendant match percentage and community match percentage.

506 One of the most intriguing findings of this experiment was that the post-deliberation differ-
507 ences in conviction rates across conditions appeared to be accounted for entirely by “confident”
508 jurors—that is, by jurors who scored above the median on a post-deliberation question that asked
509 them to rate their “ability to draw correct conclusions from numerical data, such as probabilities
510 and frequencies.” The Unconfident jurors (those below the median on this question) did not
511 differ across conditions in their conviction rates, while the Confident jurors were more likely to
512 convict in the Strong and Unknown Conditions.

513 A possible explanation for this finding is that jurors’ self-assessments of their numerical
514 abilities were accurate: the Confident jurors may have understood the statistical data better than
515 the Unconfident jurors, which would explain why they were more likely to convict when the

bullet lead evidence was highly diagnostic (Strong Condition). The Unconfident jurors, on the other hand, may have been confused by the statistical data, or uncertain about how to use it. Lacking confidence in their ability to draw correct conclusions from the bullet lead evidence, they may have elected to give it little or no weight and to rely on the other evidence in the case, which would explain their consistently low conviction rates across conditions.

On the other hand, it was also the Confident jurors who accounted for the elevated conviction rates in the Unknown Condition (where conviction rates were significantly higher than the Worthless Condition and marginally higher than the Control). Despite their asserted numerical prowess, Confident jurors may have been willing to give weight to forensic evidence in the absence of the statistical data needed to determine whether it was diagnostic. Perhaps numerical confidence is associated with trust in science and technology, and hence these Confident jurors were willing to assume in the absence of data that the bullet lead evidence had some value. Another possible explanation is that, as Beyth-Marom and Fischhoff (1983) have suggested, people (even those who are confident in their numerical skills) are better at responding rationally to data when it is presented to them than they are at dealing with problems in the absence of data.

Although intriguing, it is important to recognize that the relationship between confidence and jurors' response to the statistical data was detected during post-hoc analysis of a number of post-deliberation measures. To be certain the relationship is robust and real, and not simply adventitious, it should be confirmed through further research.

Another noteworthy aspect of this study was the effect of group deliberation on the pattern of responses across conditions. Conviction rates were lower for all conditions following deliberation, a phenomenon that has been observed in other studies (MacCoun & Kerr, 1988). Deliberation may have raised doubts about the evidence that jurors did not previously consider. It may also have raised jurors' threshold of conviction by reinforcing the standard of proof ("beyond a reasonable doubt"), making jurors more reluctant to convict on the same evidence. But deliberation did more than simply lower the conviction rate. There was a different pattern across the four conditions before deliberation than after deliberation on two of the main dependent measures: strength of case and probability of guilt. Deliberation led to decreases in strength of case and probability of guilt ratings in the Worthless, Unknown, and Control conditions, but not in the Strong Condition, thereby amplifying (or solidifying) the differences among conditions.

An obvious limitation of this study is that participants were students at a university. As a group they were younger and probably more sophisticated mathematically than the average juror. It is possible that university undergraduates are generally more confident in their ability to draw correct conclusions from numerical data than actual jurors. In light of our finding that the conviction rates of "confident" jurors were more sensitive to the diagnostic value of the forensic evidence, a difference in numerical confidence between our subject population and the actual population of jurors could be significant. For example, it might be the case that actual jurors who are confident in their numerical abilities will respond to forensic statistics in the way our Confident jurors did—showing sensitivity to diagnostic value—but that they will be greatly outnumbered by jurors who lack such confidence and who respond in the way our Unconfident jurors did. Further research exploring the effect of individual differences on people's reactions to forensic evidence would clearly be helpful for understanding how far the intriguing findings reported here can be generalized to the world at large.

The distinction between *reliability* and *diagnosticity* that was key to our analysis of the probative value of bullet lead evidence is also helpful for analyzing the value of other types of forensic evidence. To evaluate a forensic match jurors must always consider both factors. They often will encounter situations in which the diagnostic value of the "match" is reduced to some degree either because the probability of a "true match" is less than one (e.g., Finkelstein &

565 Levin, 2003; Meester & Sjerps, 2003; Evett, 1987) or because the probability of a “false match”
 566 is greater than zero, or both. Although many areas of forensic science are so poorly validated
 567 that no reliable data are available on either *reliability* or *diagnosticity*, that situation should
 568 improve in the near future as forensic scientists come under increasing pressure to improve their
 569 validation (e.g., Kennedy, 2003; Saks & Koehler, 2005). Hence, this is an opportune time to
 570 consider how these types of statistical evidence might best be presented to lay juries.

571 **Acknowledgement** The authors thank Rachel Dioso for her helpful comments on the manuscript.

572 References

- 573 Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality & Social*
 574 *Psychology*, *45*, 1185–1195.
- 575 Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*,
 576 *43*, 111–121.
- 577 Evett, I. W. (1987). On meaningful questions: A two-trace transfer problem. *Journal of the Forensic Science*
 578 *Society*, *27*, 375–381.
- 579 Faigman, D. L., & Baglioni, A. J., Jr. (1988). Bayes’ theorem in the trial process: Instructing jurors on the value
 580 of statistical evidence. *Law and Human Behavior*, *12*, 1–17.
- 581 Faigman, D., Kaye, D. H., Saks, M. J., & Sanders, J. (2002). *Modern scientific evidence: The law and science of*
 582 *expert testimony* (2 Ed.). St. Paul, MN: West Group.
- 583 Finkelstein, M. O., & Levin, B. (2003). On the probative value of evidence from a screening search. *Jurimetrics*,
 584 *43*, 265–290.
- 585 Finkelstein, M. O., & Levin, B. (2005). Compositional analysis of bullet lead as forensic evidence. *Journal of Law*
 586 *and Policy*, *13*, 119–142.
- 587 Goodman, J. (1992). Jurors’ comprehension and assessment of probabilistic evidence. *American Journal of Trial*
 588 *Advocacy*, *16*, 361–389.
- 589 Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Medicine: Communicating statistical information.
 590 *Science*, *290*(5500), 2261–2262.
- 591 Imwinkelried, E. J., & Tobin, W. A. (2003, Spring). Comparative bullet lead analysis (CBLA) evidence: Valid
 592 inference or ipse dixit? *Oklahoma City University Law Review*, 43–72.
- 593 Kennedy, D. (2003). Forensic science: Oxymoron? *Science*, *302*(5651), 1625.
- 594 Koehler, J. J., Chia, A., & Lindsey, S. (1995). The random match probability (RMP) in DNA evidence: Irrelevant
 595 and prejudicial? *Jurimetrics Journal*, *35*, 201–219.
- 596 Koehler, J. J., & Macchi, L. (2004). Thinking about low-probability events. *Psychological Science*, *15*, 540–546.
- 597 MacCoun, R. J., & Kerr, N. L. (1988). Asymmetric influence in mock jury deliberation: Jurors’ bias for leniency.
 598 *Journal of Personality and Social Psychology*, *54*, 21–33.
- 599 Meester, R., & Sjerps, M. (2003). The evidential value in the DNA database search controversy and the two-stain
 600 problem. *Biometrics*, *59*, 727–732.
- 601 Nance, D. A., & Morris, S. B. (2002). An empirical assessment of presentation formats for trace evidence with a
 602 relatively large and quantifiable random match probability. *Jurimetrics Journal*, *42*, 403–448.
- 603 Nance, D. A., & Morris, S. B. (2005, June). Juror understanding of DNA evidence: An empirical assessment of
 604 presentation formats for trace evidence with a relatively small random-match probability. *Journal of Legal*
 605 *Studies*, 395–442.
- 606 National Research Council. (2004). *Forensic analysis: Weighing bullet lead evidence*. Washington, DC: National
 607 Academy Press.
- 608 Piller, C. (2005, September 2). FBI abandons controversial bullet-matching technique. *Los Angeles Times*,
 609 pp. A-38.
- 610 Randich, E., Duerfeldt, W., McLendon, W., & Tobin, W. (2002). A metallurgical review of the interpretation of
 611 bullet lead compositional analysis. *Forensic Science International*, *127*(3), 174–191.
- 612 Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*,
 613 *309*(5736), 892–895.
- 614 Schklar, J., & Diamond, S. S. (1999). Juror reactions to DNA evidence: Errors and expectancies. *Law & Human*
 615 *Behavior*, *23*(2), 159–184.
- 616 Schum, D. A. (1994). *Evidential foundations of probabilistic reasoning*. New York: Wiley.
- 617 Schum, D. A., & DuCharme, W. M. (1971). Comments on the relationship between the impact and the reliability
 618 of evidence. *Organizational Behavior and Human Performance*, *6*(2), 111–131.

- Smith, B. C., Penrod, S. D., Otto, A. L., & Park, R. C. (1996). Jurors' use of probabilistic evidence. *Law & Human Behavior*, *20*, 49–82. 619
620
- Thompson, W. C. (2005). Analyzing the relevance and admissibility of bullet lead evidence: Did the NRC report miss the target. *Jurimetrics Journal*, *46*, 65–89. 621
622
- Thompson, W. C. (1989). Are juries competent to evaluate statistical evidence? *Law and Contemporary Problems*, *52*, 9–41. 623
624
- Thompson, W. C., & Cole, S. A. (2006) Psychological aspects of forensic identification evidence. In M. Costanzo, D. Krauss, & K. Pezdek (Eds.), *Expert psychological testimony for the courts*. Erlbaum. 625
626
- Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, *11*, 167–187. 627
628
- Thompson, W. C., Taroni, F., & Aitkin, C. G. G. (2003). How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Sciences*, *48*, 1–8. 629
630
- Tobin, W. A., & Duerfeldt, W. (2002, Fall). How probative is comparative bullet lead analysis? *Criminal Justice*, *17*, 26–34. 631
632
- United States v. Mikos*. (2003). 2003 WL 22922197, No. 02 CR 137 (ND Ill. Dec. 9, 2003). 633

UNCORRECTED PROOF

Query

Q1: Au: Please provide 3–5 keywords for this article.

UNCORRECTED PROOF