Journal of Law and Policy

Volume 13
Issue 1
SCIENCE FOR JUDGES III:
Maintaining the Integrity of Scientific Research and
Forensic Evidence in Criminal Proceedings

Article 7

2005

Compositional Analysis of Bullet Lead as Forensic Evidence

Michael O. Finkelstein

Bruce Levin

COMPOSITIONAL ANALYSIS OF BULLET LEAD AS FORENSIC EVIDENCE*

*Michael O. Finkelstein & Bruce Levin, Ph.D.***

INTRODUCTION

In crimes involving shootings, bullets may be recovered from the crime scene and unexpended cartridges found in the possession of a suspect. Did they come from the same source? The method of choice for making this determination is a compositional analysis of trace elements in the bullet lead. The Federal Bureau of Investigation (FBI or Bureau) laboratory in Quantico, Virginia, performs such analyses in federal crimes and, at the request of state law enforcement, in state crimes. It appears to be the only laboratory in the United States to do so on a regular basis.[1] Standard FBI procedure measures the concentrations of antimony, arsenic, bismuth, cadmium, copper, silver, and tin through an analytic method called inductively coupled plasma-optical

---

** Michael O. Finkelstein is a member of the New York Bar, and the adjunct faculty of Columbia Law School and University of Pennsylvania Law School. Mr. Finkelstein was a member of the National Research Council Committee on Scientific Assessment of Bullet Lead Elemental Comparison, which published a report discussed in this article. A portion of this paper was delivered by Mr. Finkelstein at the Science for Judges program held at Brooklyn Law School on March 27, 2004. Bruce Levin is a Professor of Biostatistics and Chair of the Department of Biostatistics at the Mailman School of Public Health, Columbia University.

[1] The technique has been used by crime laboratories in other countries, among them, Belgium, Canada, Germany, India, Israel, Italy, Japan, and South Africa.

emission spectroscopy. If the concentrations of all seven elements are close enough to be within measurement error, the FBI expert reports that the bullets are "analytically indistinguishable," for which we will use the shorthand term "match." The expert concludes from this that the bullets came from the same melt of lead, which is evidence that they were made at the same time by the same manufacturer, and some circumstantial evidence that they were bought at the same time by the suspect. Bullet lead evidence can thus act as a link tying a suspect to a crime.

The method for measuring these element concentrations is not in serious dispute; however, arguments frequently arise regarding the probative value of a match. Beginning in 2000, two metallurgists, one of whom had worked at the FBI, began attacking the FBI's bullet-lead matching testimony in court appearances. The Bureau responded by conducting and publishing studies on bullet-lead matching and, in 2003, retained the National Research Council (NRC) to study the matter. The NRC appointed a Committee on Scientific Assessment of Bullet Lead Elemental Comparison (Committee) to study the technique and its presentation in court. The Committee met four times between February and May of 2003, heard presentations from the FBI, bullet manufacturers, and others, made analyses of FBI data, and researched cases in which FBI experts had testified. The Committee also invited critics of the FBI to appear; one lawyer accepted the invitation, made a written submission, and addressed the committee. The principal critics, however, declined to appear or make submissions. The Committee issued its report, entitled *Forensic Analysis: Weighing Bullet Lead Evidence*, in February 2004.[2]

In discussing the Committee's findings, we will focus on two aspects of this forensic technique: (1) testimony of the FBI's expert witnesses and (2) scientific support for the reliability of the test. We begin by describing briefly the bullet manufacturing process and the FBI's procedures for analyzing the composition of bullet

---

[2] COMMITTEE ON SCIENTIFIC ASSESSMENT OF BULLET LEAD ELEMENTAL COMPOSITION COMPARISON, NATIONAL RESEARCH COUNCIL, FORENSIC ANALYSIS: WEIGHING BULLET LEAD EVIDENCE (2004) [hereinafter Report].

lead.

I. THE BULLET–LEAD MANUFACTURING PROCESS

Recycled car batteries are the principal source of the lead used in most bullets made in the United States. Secondary smelters refine the lead by melting it in kettle-type furnaces to remove or reduce the amounts of certain trace elements. Manufacturers specify antimony within a range and other trace elements merely as maximum amounts. For example, one manufacturer specified a range of 0.85% ± 0.15% antimony for lead used for .22 long rifle bullets; maximal impurities for the other elements ranged from 0.2% for tin to 0.01% for silver.[3]

The kettles have a capacity of 25 to 150 tons a day. Molten lead is poured from the kettle to cast sows, billets, pigs or ingots. A sow will weigh 2,000 lbs; a billet, 70 to 350 lbs; and a pig or ingot, 60 to 125 lbs. Castings by secondary smelters are sold to bullet manufacturers, who sometimes remelt and recast the lead. The cast lead is made into a wire by squeezing it through a narrow opening, like toothpaste from a tube, in a machine called an extruder. The wire is then cut into slugs, from which the bullets are shaped. The bullets are stored in a bin until they are assembled into cartridges. Slugs from several melts may be placed in the same bin. The number of bullets that are made from a melt or pour will vary widely. For example, a melt pot of 200,000 lbs will yield 35,000,000 .22-caliber bullets, while a pig or ingot will yield 10,000 to 20,000 bullets.[4] The yield for larger caliber bullets will be smaller. When cartridges are packed into boxes (which typically hold fifty bullets each) there is routinely more than one melt represented in a box. Bullets from a single melt may be represented in thousands of boxes.

II. THE FBI'S ANALYTIC METHOD AND MATCHING CRITERIA

In making its analysis, the Bureau takes three fragments from

---

[3] *Id.* at 78, Table 4.4.

[4] *Id.* at 74, Table 4.3.

each bullet, measures the seven elements for each fragment, and then takes the average of the replicate measurements for each element.[5] It also estimates the standard deviation of its measurements from the variation in measurements of the three fragments. The concentrations in each fragment depend on the element involved. For example, in one sample they ranged from 1.13% by weight for antimony to 0.0046% for silver.[6] The standard deviation of the measurement ranged between 0.06 percentage points for antimony to 0.0001 percentage points for silver.[7]

The FBI currently uses one of two criteria to determine whether two bullets match. In the first method, the Bureau calculates a confidence interval of two standard deviations on either side of the average measurement for each element.[8] If the confidence intervals for the crime scene bullets and the suspect's bullets overlap to any extent for all seven elements, the bullets are declared to match. Because the standard deviations for the measurements of two bullets tend to be the same for the same element, differences of up to four standard deviations between measurements effectively are allowed for a match.

This is not the standard way to define a match window in statistical science but rather an approximation of it. The usual criterion would define a window by the standard error of the difference between the two measurements, which is the square root of the sum of the variances for the two average measurements. If

---

[5] The Bureau in fact makes three measurements for each fragment and takes the average as the measurement for that fragment. Since the committee treated that average as a single measurement for each fragment, we shall do so here. *Id.* at 29 n.5.

[6] Robert D. Koons & Diana M. Grant, *Compositional Variation in Bullet Lead Manufacture*, 47 J. FORENSIC SCI. No. 5 1, 3 Table 8 (2002). This sample may understate the full range of variation; most bullets in the 1,837 sample contain one or more elements at concentrations more like 0.0002% at the low end. *See* discussion *infra* note 42.

[7] Antimony is present in larger quantities because it is added to the lead to increase hardness; except for maximal limitations, the others are uncontrolled trace elemental impurities.

[8] When we refer to standard deviation, we mean the standard deviation of the three measurements, not the standard error of the mean measurement.

the usual 5% test of statistical significance is applied as the criterion for the test as a whole, it must be adjusted for multiple comparisons when dealing with each element. In the commonly used Bonferroni method of adjustment, because there are seven chances to reject the null hypothesis (one for each element), the level of significance required to reject the null hypothesis for each element would be 0.05/7 = 0.00714.[9] Applying a standard statistical test, this probability is reached at 5.07 standard errors.[10] Assuming that the standard deviation of measurement (denoted by σ) is the same for both bullets, the bullets would be declared to match if the difference between them is no more than 5.07 x $(2/3)^{½}\sigma = 4.14\sigma$.[11] On this assumption, the FBI's 4σ match window is even slightly narrower than statistical convention would dictate. The size of the match window is important because the wider the window, the greater the number of declared matches and, with that, a greater risk that bullets will be declared matches even though they come from different melts.

This does not mean that the Bureau's two-standard-deviation rule is beyond criticism. From the criminal justice point of view, what matters most is that the test has sufficient statistical power. That is, if bullets come from different melts, the test with high probability will reject the null hypothesis that they are from the same melt. The practice of estimating standard deviations from only three observations significantly increases the size of the required match window, thus reducing the power of the test. It is for this reason that the Committee recommended that standard

---

[9] With this adjustment, the level of significance for the overall test would be no greater than 5%. For a discussion of Bonferroni and other methods of adjustment, see MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 211 (2d ed. 2001).

[10] The test is a *t*-test with four degrees of freedom. For a discussion of the *t*-test, see *id.* at 222.

[11] If the variance of a single measurement is σ^2, then the variance of the average of the three measurements is $\sigma^2/3$; the variance of the difference between the two average measurements (assuming σ^2 is the same for both bullets) is $\sigma^2/3 + \sigma^2/3$, with standard error equal to $[\sigma^2/3 + \sigma^2/3]^{1/2}$. A 95% two-sided confidence interval, making the Bonferroni adjustment for the seven comparisons, is equal to 5.07 standard errors, or 5.07 x $(2/3)^{½}\sigma = 4.14\sigma$.

deviations be estimated from historical data.[12] If this were done, the number of standard errors required for the difference between the two measurements would be reduced to about 2.2σ, or almost half of the 4σ window.[13] This would significantly increase the power of the test.

In the second method, the FBI declares a match when the ranges of the two average measurements for an element overlap, that is, the largest of the three measurements for one bullet is greater than the smallest measurement for the other bullet. The range window is generally narrower than the two-standard-deviation window and would tend to generate fewer false positives. Range overlap methods are appealing for statistical reasons because they are non-parametric and, therefore, universally valid for any distribution of elemental compositions. They are also appealing for the practical reason that they are easier to explain to a jury.

III. THE FBI'S TESTIMONY

Connecting a bullet from a crime scene to a particular suspect involves two steps. First, the FBI must perform a bullet lead analysis to determine whether the bullets found at the crime scene came from the same melt as the bullets found in the possession of the suspect. Second, an inference must then be made that the matching bullets came from the defendant. Problems arise regarding both of these steps.

---

[12] Report, *supra* note 2, at 69. The committee's recommendation assumes that good estimates of the standard errors can be made from the historical data; there are difficulties in this assumption and the idea has not been tested.

[13] Using a relatively large amount of historical data to estimate standard errors would justify using the normal distribution, which for the Bonferroni-adjusted significance probability of 0.00714 is 2.69 standard errors. $2.69 \times (2/3)^{1/2}\ \sigma = 2.2\sigma$. Power would be further increased if methods other than Bonferroni were used to adjust for multiple comparisons. One such alternative is the Hochberg method. *See* Yosef Hochberg, *A Sharper Bonferroni Procedure for Multiple Tests of Significance*, 75 BIOMETRIKA 800 (1988); FINKELSTEIN & LEVIN, *supra* note 9.

A. Bullet Analysis

In some cases, FBI witnesses have testified in a way that was considered unobjectionable by the Committee; for example, witnesses have testified that bullets were analytically indistinguishable,[14] or that the bullets *could have* come from the same melt or source lead.[15] In a number of cases, however, the Committee found that FBI testimony had gone considerably beyond what the science would justify. FBI witnesses have testified without qualification that matching bullets came from the same melt or source of lead, or were manufactured by the same company at the same time. Witnesses also have testified and prosecutors have argued, with various qualifications (or none), that bullets came from the same box of cartridges.

For example, in *United States v. Davis*, the expert testified unequivocally that the "bullets must have been manufactured at the same Remington factory, must have come from the same batch of lead, must have been packaged on or about the same day, and could have come from the same box."[16] In *State v. Washington*, the FBI expert testified that the elemental profile of bullets in a melt was "unique."[17]

Further, in *State v. Noel*, the prosecution in summation referred to snowflakes and fingerprints to describe the uniqueness of melts of lead, and argued that the FBI witness' testimony was reliable scientific evidence not only that the bullets "came from the same source of lead at the manufacturer," but also that they were "sold

---

[14] *See, e.g.*, Wilkerson v. State, 776 A.2d 685, 689 (Md. Ct. Spec. App. 2001) (quoting a witness as testifying that "[t]he lead material in one bullet and one projectile was analytically indistinguishable, as was the lead in one bullet and the other two projectiles").

[15] *See, e.g.*, State v. Krummacher, 523 P.2d 1009, 1012 (Or. 1974) (stating that that the "analyses showed that the bullet could have come from the same batch of metal as the group of bullets which was taken from defendant's home but not from the same batch as any of the other groups").

[16] 103 F.3d 660, 673-74 (8th Cir. 1996).

[17] Trial Transcript of Testimony of Charles Peters at 9, 21-22, State v. Washington, No. 96-GS-40-10316 (S.C. Ct. 1998).

in the same box."[18] The New Jersey intermediate appellate court reversed the conviction, holding that "the clear import of the fingerprint and snowflake comparison was to suggest to the jury a scientific certainty in the inference that defendant had possessed both bullets and to suggest to the jury a conclusiveness of that inference that clearly was not warranted."[19] On further appeal, however, the New Jersey Supreme Court conceded that the prosecutor's argument may have been "excessive," but held that it might pass as "fair comment" and reinstated the conviction.[20]

More acceptably, witnesses have testified that the laboratory finding that bullets matched was "consistent with their having come from the same box of ammunition,"[21] or was evidence that they "could have come from the same box or another box manufactured on the same day."[22] Although literally true, these formulations invite the jury to overestimate the evidence. Nevertheless, they have been upheld on appeal.[23]

The Committee disapproved of such testimony and argument. The obvious point is that witnesses who testify without qualification and prosecutors who argue in this manner do not admit that melts are not unique in elemental composition and that tests have measurement error. Thus, the fact that bullets match is only a probability statement that they came from the same melt. The Committee also disapproved of testimony suggesting that matching bullets came from the same box because of the large number of boxes that are filled, wholly or partly, with bullets from

---

[18] 723 A.2d 602, 608 (N.J. Sup. Ct. 1999).

[19] State v. Noel, 697 A.2d 157, 165 (N.J. Super. Ct. App. Div. 1997), *rev'd*, 723 A.2d 602 (N.J. Sup. Ct. 1999).

[20] *Noel,* 723 A.2d at 607.

[21] State v. Reynolds, 297 S.E.2d 532, 534 (N.C. 1982) ("Further, neutron activation analysis revealed that the bullets taken from Morgan and Stone and the ammunition found with the defendant were of the same chemical composition, consistent with their having come from the same box of ammunition.").

[22] State v. Grube, 883 P.2d 1069, 1078 (Idaho 1994).

[23] *See, e.g.*, United States v. Davis, 103 F.3d 660, 667 (8th Cir. 1996) (holding that "[t]he expert testimony demonstrated a high probability that the bullets spent at the first robbery and the last robbery originated from the same box of cartridges").

a single melt. But the Committee's objections also went deeper than this: they included the case in which a witness testifies that matching bullets *probably* came from the same melt. To understand the Committee's objection to such testimony, one has to consider the probative force of the matching evidence in more detail.

The probative force of finding that two bullets match can be expressed by the odds that they came from the same melt, given that they match. These odds are usually called posterior odds because they are posterior to, or conditioned upon, the matching evidence. By Bayes's theorem of elementary probability, these odds are equal to the prior odds multiplied by the likelihood ratio. The equation can be written as follows:

$$Posterior\ odds = Prior\ odds \times Likelihood\ ratio$$

The terms of the equation can be defined as:

$$\frac{P[same\ melt \mid match]}{P[different\ melts \mid match]} = \frac{P[same\ melt]}{P[different\ melts]} \times \frac{P[match \mid same\ melt]}{P[match \mid different\ melts]}.$$

In this equation, $P[same\ melt \mid match]$ is read as the conditional probability that two bullets would have come from the same melt, given that they match, with analogous interpretations for the other conditional probability terms.

Unlike the other terms in the equation, the prior odds are not based on conditional probabilities, but rather are the odds that one would give that the bullets came from the same melt before they are tested for match status. Prior odds must be based on the other evidence in the case that connects the suspect to the crime. Typically, there is no objective way of estimating prior odds; this must be done subjectively, based on the odds that the factfinder would give in betting on the proposition, in this case, that the bullets came from the same melt (which the factfinder might explicitly or implicitly infer from other evidence of the suspect's guilt). Expert witnesses have no particular expertise beyond that of the lay factfinders to estimate prior odds and, consequently, have

no basis for using *their* estimates of prior odds to testify as experts to posterior odds based on the matching test.

The second part of the equation, the likelihood ratio, is a measure of the strength of the matching evidence. In theory, this can be determined on an objective basis and is an appropriate subject of expertise. In this context, the likelihood ratio for the matching evidence is equal to the probability of observing a match, given that the bullets came from the same melt, divided by the probability of observing such a match if the bullets came from different melts. The larger the likelihood ratio, the stronger the evidence.

Consider first the numerator of the likelihood ratio. If melts of lead were perfectly homogeneous in elemental composition, the probability of observing a match if two bullets came from the same melt would be one, so the numerator of the likelihood ratio would be one.[24] Allowing 5% for measurement error, for example, the numerator would be 0.95.

The denominator of the likelihood ratio is the probability of declaring a match when two bullets come from different melts. This type of error is known as a false positive. Assume that the rate of false positives is 1/1000. In that case, the denominator of the likelihood ratio would be 1/1000, and the likelihood ratio, assuming the numerator to be 0.95, would be 0.95/(1/1000) = 950. On these assumptions, matching is strong evidence that the bullets come from the same melt.

If the assumed value for the likelihood ratio is fully supported (an important subject that we will address), what can the expert say in such a case? She cannot testify to the odds or probabilities that the bullets came from the same melt because those odds depend on the prior odds and, as noted, she has no more expertise in appraising those odds than the lay factfinders. Moreover, the expert's prior odds may not be the same as the jury's and so her testimony would not "fit" the case.[25] What the expert can describe is either the likelihood ratio associated with a match or its effect on

---

[24] This assumes that the test was error free.

[25] *See, e.g.*, State v. Spann, 617 A.2d 247 (N.J. Super. Ct. 1993) (rejecting an expert's posterior probability testimony based on the expert's assumed 50% prior probability).

the posterior odds. The description of the likelihood ratio would be as follows: The probability of seeing a match is 950 times greater if bullets came from the same melt than if they came from different melts. Alternatively, the expert may describe the effect on the posterior odds by stating: The odds that two bullets came from the same melt are increased 950 times when a match is identified. Notice that in both formulations, the witness does not give the absolute level of probabilities or odds, for example, that the bullets probably came from the same melt, but only the change in probabilities or odds attendant on finding a match.

In light of the potentially misleading nature of bullet lead evidence, the Committee recommended that FBI experts limit their testimony to a description of the likelihood ratio associated with a given match, as described above. In fact, the Committee went further and recommended that, in view of the uncertainties in the estimates of the error rates, FBI experts should not quantify the likelihood ratio, but should only say that bullets from the same melt are more likely to be analytically indistinguishable than bullets from different melts, or that the fact that bullets are analytically indistinguishable makes it more likely that they came from the same melt.[26] The testimonial limitations suggested in the Committee's recommendation obviously would apply not only to bullet lead, but to all kinds of expert identification testimony. If generally adopted, they would change the way opinions in expert testimony are expressed in our courts.

[26] Report, *supra* note 2, at 107.

B. Defendant as the Source

When confronted with FBI testimony that bullets match, defense counsel commonly and correctly point out that many bullets are made from a single melt. The number ranges from thousands to millions, and there is usually no way of knowing how large the melt was for the bullets involved in a case. Moreover, geographical distribution of bullets from a melt may affect the probability of finding matching bullets. For example, if bullets made from a single melt are sent to a distributor in a small town or to a vendor in a city neighborhood, every gun owner in the area may have them. But distribution patterns, as the Committee's report makes clear, are not available to the public and were not made available by manufacturers to the Committee.[27]

These gaps in our knowledge do not affect the likelihood ratio associated with bullet lead matching or the validity of test. Whether there are many or few bullets made from a melt, or how they are distributed, does not affect the probability that the bullets would match if they came from the same melt or the probability that they would match if they came from different melts. What these factors do affect is the second inference, that is, the probability that the crime-scene bullet came from the defendant if the bullets came from the same melt. While it is reasonable to conclude that it is more likely that the crime-scene bullet came from the defendant if it matches the defendant's other bullets than if it did not, no FBI bullet-lead witness has the expertise to attach a probability to that inference. Thus, the report ruled out any testimony as to that probability *as a matter of expertise*, recognizing that jurors would have to take that step, if at all, on their own.[28]

The difficulty is that jurors have no common experience by which to judge the strength of the connection between a bullet's coming from the same melt and from the same person. Bullet-lead matching has been described as circumstantial evidence akin to testimony that the suspect wore a red ski parka coupled with an

---

[27] *Id.* at 102.

[28] *Id.*

investigator's finding a red ski parka in the suspect's closet. But in that case, jurors may have some feeling (possibly quite wrong) for the frequency of red parkas and their distribution, while in bullet lead matching they have no experience with either subject. The absence of expert testimony on the subject of frequency and distribution significantly weakens the value of this evidence.

The effect of the size of the melt and the distribution of bullets made from it can be formalized by considering the likelihood ratio (LR) for a match, given that the crime-scene bullet came from the defendant or from someone else. As previously noted, given that the bullets came from the same melt or from different melts, the conditional probabilities of a match do not depend on whether or not the defendant is the source of the bullet. It follows that the LR can be expressed as:

$$\mathrm{LR} = \frac{P[M \mid G]}{P[M \mid \overline{G}]} = \frac{P[M \mid S] \cdot P[S \mid G] + P[M \mid \overline{S}] \cdot P[\overline{S} \mid G]}{P[M \mid S] \cdot P[S \mid \overline{G}] + P[M \mid \overline{S}] \cdot P[\overline{S} \mid \overline{G}]},$$

where M is the event that the bullets match; S is that they come from the same melt; $\overline{S}$ is that they come from different melts; G is that the crime-scene bullet came from the defendant; and $\overline{G}$ is that it came from someone else. Assuming that false positives are rare, $P[M \mid \overline{S}]$ will be quite small and the second term of the numerator and denominator may be disregarded. Moreover, because the FBI will test all of the defendant's bullets, it is very likely that if the defendant is the source of the crime-scene bullet, the Bureau will find at least one of defendant's bullets from the same melt; thus $P[S \mid G]$ will be close to 1. With these reductions, LR will be approximately equal to $1/P[S \mid \overline{G}]$, the reciprocal of the probability that an innocent person would have a bullet from the same melt as the crime-scene bullet. This probability depends on the size of the melt and distribution practices, as previously noted.

IV. THE ISSUE OF SCIENTIFIC SUPPORT

A. Homogeneity of Melts

The validity of even properly circumscribed testimony depends on scientific support for the likelihood ratio associated with the test. Assumed values for both the numerator and denominator of the ratio have been subjected to criticism. The numerator is the probability that the FBI would declare a match if the two bullets came from the same melt. As noted, this turns on the homogeneity of melts and, to a lesser extent, on measurement error. If melts are inhomogeneous, then the probability of finding a match will be reduced for bullets from the same melt.

The Committee spent considerable time examining the evidence bearing on homogeneity. Experts agreed that there was some inhomogeneity in a melt, but disagreed on its extent and significance.[29] The melting process causes convective stirring of the molten lead in the kettle, so that all of the bullets made from the same melt would tend to have the same elemental composition. However, some smelters add lead to replenish the pot during a pour and this can change the composition of billets created during an extended pour. Inhomogeneity may also arise because solutes tend to migrate to the center of the billet during solidification. On the other hand, the extrusion process is thought to reduce inhomogeneity caused by segregation because the flow of the solid is turbulent as the billet is squeezed through the mouth of the die.[30]

After considering conflicting reports on homogeneity, the Committee concluded that it was "unclear whether macro- and microscale inhomogeneities are present at some or all of the stages

---

[29] *Compare* Robert D. Koons & Diana M. Grant, *Compositional Variation in Bullet Lead Manufacture*, 47 J. FOREN. SCI. 950 (2002) (finding no significant inhomogeneity), *with* Erik Randich et al., *A Metallurgical Review of the Interpretation of Bullet Lead Compositional Analysis*, 127 FOREN. SCI. INT'L. 174 (2000) (finding significant inhomogeneity).

[30] Report, *supra* note 2, at 82-84.

of lead and bullet production and if such inhomogeneities would affect [compositional analysis of bullet lead or] CABL."[31]

Arguably, it is unnecessary to resolve the issue of homogeneity to make use of CABL. In its report, the Committee dealt with the problem by making the reasonable assumption that at least some part of a melt had to be homogeneous within measurement error (otherwise there would be no matches) and by treating that part, in effect, as a smaller melt. The Committee called this a "compositionally indistinguishable volume of lead" or "CIVL."[32] The CIVL might be as large as a vat of molten lead, the composition of which was not altered during the pouring of bullets, or a series of billets that were poured before the composition of the vat was altered by the addition of more lead to replenish the vat. The report observed that the probative force of a match would be increased if only a part of a melt were homogeneous because fewer matching bullets would be made from a smaller volume of lead.[33]

B. Measurement Error

In appraising the number of bullets produced by the CIVL, however, the report recognized that it would have to be assumed, conservatively, that the entire melt was involved because the size of the CIVL could not be known.[34] With this assumption, the numerator of the likelihood ratio would be one, except for measurement error. Measurement error could potentially affect the test by causing it to declare that bullets do not match when, in fact, the bullets came from the same melt. This type of error makes the numerator of the likelihood ratio less than one and reduces the probative force of the test, other things being equal. However, as noted above, the effect would be small. Moreover, by declaring a non-match when bullets in fact come from the same melt, the

[31] *Id.* at 82.
[32] *Id.* at 98.
[33] *Id.*
[34] *Id.*

effect of this error will generally be exonerating, which makes it of lesser concern than a false positive.[35]

More significantly, the allowance for measurement error creates the match window that leads an FBI expert to conclude that two bullets that have different measurements, but are within the window, are "analytically indistinguishable." The size of the window affects the denominator of the likelihood ratio, and it is the denominator that is of greater concern because it measures the rate of falsely incriminating evidence. The value of the denominator, assumed to be very small, is what makes the likelihood ratio so large.

C. The Problem of False Positives

Scientific support for the assumed low rate of false positives raises a controversial issue.[36] The risk of coincidental matches is the point of focus for critics of the FBI, who argue that the risk of false positives is much higher than FBI witnesses admit.[37] On the stand, FBI experts have gone so far as to describe the rate of false positives as zero.[38] In response to claims that the false positive rate was high, the FBI made a study of the matter and the question now is whether that study justifies the conclusion that the FBI would draw from it.[39]

---

[35] But not always. For example, if one of two suspects did a shooting, finding that the bullets did not match those in the house of one suspect would be some evidence, perhaps not very strong, pointing to the other.

[36] It might be thought that if the FBI testifies only that the bullets are analytically indistinguishable that there is no false positive because the statement is true when there is a coincidental match. However, what follows from that conclusion, stated or unstated, is that the bullets come from the same melt, and that is false.

[37] Randich et al., *supra* note 29.

[38] *See, e.g.*, Trial Transcript of Testimony of Charles Peters at 9, 21-22, State v. Washington, No. 96-GS-40-10316 (S.C. Ct. 1998).

[39] A weakness of the study is that it has not been published. Robert Koons, the principal FBI architect of the study, is in the course of preparing it for publication. The description of the study, which follows, is based on presentations made by Mr. Koons to the commission. *See* Report, *supra* note 2, at 28-35, 54-60.

In conducting the study, the FBI used its archive data on approximately 23,000 bullets that were collected over fourteen years in about 1,000 cases. From this inventory, the Bureau collected data for as many bullets as it could reasonably be sure came from different melts. To arrive at this sample, the Bureau took each case in its inventory and selected from it data for one bullet of each type (principally caliber), of each alloy class (defined by the amount of antimony in the lead), and from each manufacturer (when this was known). If there were multiple bullets in the same cluster, the Bureau picked one at random. The Bureau also included data for some bullets that were not connected with any case. However, it did not make cross–case comparisons, that is, it did not rule out multiple bullets from the same alloy class of the same manufacturer across cases. Thus, some bullets included in the sample might have come from the same melt.

The Bureau collected data from 1,837 bullets as a result of this winnowing process.[40] It then looked at every possible pair of bullets and counted the number of matches by its two-standard-deviation criterion. There were 1,686,366 pairs, of which 693 matched, or about 1 in 2,433. Because the bullets in the selected sample came primarily (if not exclusively) from different melts, the results of the study indicate that the probability of a false positive would be less than 1 in 2,433. Although current practice is to measure concentrations of seven elements, not all bullets in the selected sample had all seven elements measured. The rate of false positives is even lower if attention is confined to those bullets in the selected sample for which all seven elements were measured. There were 854 such bullets and 47 pairs matched by the two-standard-deviation criterion.[41] With 364,231 possible pairs and 47 matches, the rate of false matching is less than 1 in 7,750. These results strongly support the FBI's opinion that a positive test is highly probative evidence that matching bullets came from the same melt.[42]

---

[40] We refer to this process and the collection of bullets as the "selected sample."

[41] Report, *supra* note 2, at 190, Table K.8.

[42] The FBI frequently compares one or more crime-scene bullets against multiple bullets found in the possession of the suspect. In such cases involving

The Committee's report faced two ways on the rate of false positives. On one hand, the report determined that "CABL is sufficiently reliable to support testimony that bullets from the same compositionally indistinguishable volume of lead (CIVL) are more likely to be analytically indistinguishable than bullets from different CIVLs."[43] Although this testimony could be given even with a high rate of false positives,[44] the concept of *reliable* scientific evidence surely implies that the procedure's rate of false positives will be low. This assumption is made explicit in other places in the report. For example, the report sets out a proposed "boiler plate" informational sheet to be distributed to lawyers and judges with the FBI's laboratory report. The Committee suggests that the sheet include the following: "Considering the thousands of 'batches' of lead produced over a number of years, there is a reasonably high probability that some will repeat. However, the probability that any given composition would repeat within the next several years could be expected to be quite low."[45] This sounds reasonable, but neither the report nor the sheet supports either statement, nor defines the phrase "quite low."

One might think that these findings were based on the FBI study, even though it is not cited in support of them. However, although the report confirms the FBI's result, it puts the study to one side on the ground that the sample collected by the FBI was not a random sample and could have been biased.[46] The possibility of bias is based on the fact that the winnowing process could have made the bullets in the selected sample further apart in elemental

---

multiple comparisons, the usual statistical practice is to multiply the probability of a false positive in a single comparison by the number of comparisons to obtain an upper bound for the probability of one or more false positives among the multiple comparisons. Thus, if many comparisons are made, the risk of a false positive can become much larger than when a single comparison is made. *See* FINKELSTEIN & LEVIN, *supra* note 9, at 59-60.

[43] Report, *supra* note 2, at 107. This finding is repeated in Chapter 5, Major Findings and Recommendations. *Id.* at 112.

[44] For example, if the numerator of the likelihood ratio were 1, the described testimony would be literally true if the rate of false positives were any number less than 100%.

[45] *Id.* at 168.

[46] *Id*. at 40, 49, 55.

composition than bullets in some relevant population, producing a lower rate of false matches in the sample than in the population.[47]

Instead of the low rates of false positives generated in the FBI's study, the authors of Chapter 3 of the Committee's report estimated remarkably higher rates of false positives using computer simulations.[48] However, these simulated rates are unlikely to be good estimates of the actual rates of false positives. The simulations were of probabilities *conditional* on various assumed differences in elemental concentrations between the crime scene bullets and the suspect's bullets; there was no attempt to calculate an unconditional probability that would incorporate the likelihood that differences as small as those assumed would in fact be encountered in bullets from different melts. Taking those likelihoods into account would probably decrease the simulated rates.[49] This is a striking omission given that the risk of false positives is the principal point of attack by critics of the FBI.

D. Testimony of Bullet Lead Analysis and Daubert

Thus, the FBI study is the only current basis for estimating the unconditional rate of false positives. But since that study was rejected for possible bias, the Committee's conclusion that the evidence was reliable, while not unreasonable, would seem to lack

---

[47] *Id.* The committee did not specify a relevant population and referred to it somewhat inconsistently. *Compare id.* at 56 ("all bullets collected by the FBI in criminal investigations"), with *id.* at 49 ("a full subset of bullets drawn from different melts"). We suggest that a useful population would be all bullets sent to the FBI *that came from different melts*.

[48] *See id.* at 59, Table 3.10 (false positive rates between 12.7% and 40.4%).

[49] In the simulations, the assumed differences between bullets would appear to be smaller than the average actual differences between bullets from different melts, which increases the simulated rate of false positives. Table 3.10 reports, in columns, simulated false positive rates for differences in elemental composition between bullets assuming a range of 3% to 10%, whereas the actual differences for five of the elements in the selected sample range between 20% and 52%. The footnote to the table acknowledges that "the columns represent differences in bullets that are relatively small given the distribution of between-bullet differences from the 1,837 bullet set. One would expect the false match probability to be smaller for larger differences between bullets." *Id.* at 59.

scientific support (at least none is cited within the report). The existence of such support is required for admissibility of expert scientific testimony under *Daubert v. Merrill Dow Pharmaceuticals*.[50]

The FBI has argued that its extensive experience is a reasonable substitute for a quantitative study and in some courts it may be seen as such. In fact, Federal Rule of Evidence 702 has long affirmed that an expert may be qualified on the basis of "experience" to testify to "technical or other specialized knowledge."[51]

The rationale for *Daubert*'s strict standards suggests, however, that bullet lead analysis is not a field in which one may rely on experience in the absence of a rigorous study. One of the factors listed by the court in *Daubert* as the test of scientific knowledge is the known or potential error rate.[52] The FBI's extensive experience does not give it a solid basis for determining the rate of false positives because the examiners have no way of knowing whether the bullets they matched in fact came from the same melt. As a result, the FBI's testimony is delivered with the impressive credentials of hard science, but the core of the matter is a distinctly unscientific leap of faith. Jurors (and some judges) are apt to ascribe at least some of the authority of the scientific part of the enterprise to the expert's unsupported belief—delivered with equal conviction—that the technique she uses has a low error rate.

Of course, the belief of the expert witness is unsupported only if one does not credit the FBI's study. In *United States v. Mikos*,[53] the only opinion that directly addresses the validity of the study under *Daubert*, District Judge Guzman rejected the study and excluded the expert's proffered opinion that the bullets probably came from the same melt.[54] Judge Guzman based his ruling on the

---

[50] 509 U.S. 579 (1993) (holding that the general acceptance of a scientific technique is not a precondition for admission of expert testimony based upon that technique so long as the standards of reliability and relevance under the Federal Rules of Evidence are met).

[51] FED. R. EVID. 702.

[52] *Daubert*, 509 U.S. at 594.

[53] 2003 WL 22922197 (N.D. Ill. 2003).

[54] *Id.* at *6.

conclusion that the FBI's sample was not randomly selected. The court wrote that it agreed with the defense's argument that the FBI's collection methods lacked "any scientifically accepted sampling procedure," and concluded that because "the FBI's historical database fails to satisfy accepted scientific methodology . . . [it] cannot form the basis for expert opinion testimony under *Daubert.*"[55]

V. A NEW APPROACH FOR THE FBI

The FBI is presently considering how to respond to the *Mikos* case. The court in *Mikos* may have been too quick in rejecting the Bureau's selected sample study as unscientific. Because most of the cases in the FBI's inventory were included, the FBI's study did not involve a sample of cases, but rather a sampling of bullets from clusters. Multi-level sampling, including cluster sampling, is a well-accepted statistical technique.[56] Nevertheless, the reaction of the

---

[55] *Id*. at *5. The court also commented that the FBI's sample of 1,837 bullets was too small to be reliable to extrapolate to 150 billion bullets made in the United States during the past thirty years and that there was no precise and generally accepted definition of a "source" or "batch." *Id*.

These objections do not seem to be valid. The fact that sources may come in different sizes and may be inhomogeneous does not, for the reasons already given, detract from the evidence and may in fact add to it when a match is found. The fact that the FBI inventory is small relative to the universe of bullets is not of concern for three reasons: (1) the population to be represented arguably is the population of bullets sent to the FBI that are from different melts and this population is not huge relative to the sample; (2) in any event, the size of the population being represented is irrelevant to the scientific acceptability of the sample; and (3) if the population were more diverse than the sample, as the court implies, the study would overstate, not understate, the rate of false positives.

A more telling objection would be that the FBI's inventory includes bullets analyzed over the past 14 years. In that time elemental composition may have changed, creating larger differences between bullets made over time than across those made currently. For example, the Bureau has noted a decline in silver during this period. The influence of this factor has yet to be studied. The Bureau acknowledges this point, but replies that sometimes old bullets are found in current cases, so one cannot assume that all comparisons will be made between currently made bullets.

[56] *See, e.g.*, FINKELSTEIN & LEVIN, *supra* note 9, at 257-60.

Mikos court suggests that it will be very difficult, if not impossible, for the Bureau to winnow out bullets from its inventory to ensure that they come from different melts without incurring the risk that the sample so selected will be rejected as possibly biased. We therefore suggest that the Bureau take a different tack and select a random sample of bullets without eliminations for same-melt status and compute the rate of positives, both true and false, as an upper bound for the rate of false positives. Our preliminary analysis indicates that such positives would be quite rare—less than one in a thousand—which would seem to provide ample support for the reliability of CABL and justify the Committee's conclusion in that regard.

In our analysis, a cluster of bullets is a group of bullets of the same characteristics (for example, alloy class, style, etc.) organized such that only one of them would be picked under the FBI's selected sample procedure. We make the "worst case" assumption that the clusters are perfectly homogeneous, that is, that all the bullets in a cluster would match each other. It follows that when two bullets from two different clusters match, all the bullets in the two clusters would match each other and when two bullets from different clusters do not match, then none of the bullets in the two clusters would match across the clusters.[57]

Let there be n cases in the FBI's selected sample and assume that there are two clusters of bullets in each case; thus there are $2n$ clusters. Assume further that there are m matching pairs. Under the FBI's method of taking one bullet from each cluster, the rate of false matches would be

$$R = \frac{m}{\binom{2n}{2}}.$$[58]

[57] This applies only across clusters; they would still match within the clusters.

[58] The expression $\binom{2n}{2}$ should be read as "2n choose 2," the number of

Now assume that each cluster is of size *c* and that all bullets are included in the study, not merely one from each cluster. What is the rate of matches (say, *R'*), both true and false? On the "worst case" assumption previously described, the number of matches from the *m* matching pairs of clusters is $m\binom{2c}{2}$ and the number of matches from the *2n-2m* non-matching clusters is $2(n-m)\binom{c}{2}$. Because the total number of pairs of bullets is $\binom{2cn}{2}$, the rate of matching is $R' = \dfrac{m\binom{2c}{2} + 2(n-m)\binom{c}{2}}{\binom{2cn}{2}}$.

Using the simplifications $\binom{2n}{2} = \dfrac{2n(2n-1)}{2} \approx 2n^2$ and $\binom{2cn}{2} = \dfrac{(2cn)(2cn-1)}{2} \approx 2n^2c^2$, the above expressions for *R* and *R'* become $R \approx \dfrac{m}{2n^2}$ and $R' \approx \dfrac{m + n(1-\frac{1}{c})}{2n^2}$. Thus, going from *R* to *R'* adds only about $\dfrac{n(1-\frac{1}{c})}{2n^2} = \dfrac{c-1}{2nc}$ to the rate of matching. When *n* is large, this is not a large addition and it leaves the rate of matching quite low.

For example, suppose that there are *n* = 1,000 cases (as the FBI has in its inventory), each with two clusters (about the average number per case in the FBI selected sample study) and that each cluster has *c* = 10 bullets (the average size for the FBI clusters in its study). Suppose further that the rate of matching in the selected

---

pairs that can be chosen from 2*n* bullets.

sample study is R=1/2400, as the FBI found. Consequently, $R' \approx R + \frac{9}{20{,}000} = \frac{1}{1{,}154}$.

This last result is an upper bound for the rate of matches, both true and false, and is thus a kind of upper-upper bound for the rate of false matches.[59] Of course, this is only an estimate using our worst case assumption and an average size for clusters; the FBI would have to draw a random sample and apply its criteria to demonstrate convincingly that the rate of matching is low. To add credibility, any new study should be designed, or at least approved, by outside experts, and the results should be published in a peer-reviewed journal.

[59] The rate would be even lower if attention is confined to bullets for which all seven elements have been measured. *See supra* text accompanying note 41.