

---

*Are bullets found at crime scenes traceable enough to be admissible in court?*

---

# Data Integrity and the Scientific Method: the Case of Bullet Lead Data as Forensic Evidence

Clifford H. Spiegelman and Karen Kafadar

**Editor's Note:** *The authors were members of the NRC Committee on Scientific Assessment of Bullet Lead Elemental Compositional Comparison, which formally ended in February 2004. The opinions expressed in this article are those of the authors and do not represent the views of this committee, except when quoted directly from the report. The authors are grateful to W. Tobin, a former FBI metallurgist, for information about measurement protocol and for Figures 3 and 4.*

Forensic evidence can be a critical component of courtroom proceedings in assessing the guilt or innocence of a defendant, particularly in the absence of eyewitnesses or irrefutable evidence. In forensic analysis, experts try to assess whether evidence found at the crime scene (CS) is consistent with, or 'matches,' the evidence found in the possession of a potential suspect (PS). Some examples of such evidence include blood stains, human hair, and fragments of material such as glass, paint, or bullets.

Federal Rules of Evidence govern the admissibility standards for such evidence in federal courtrooms as well

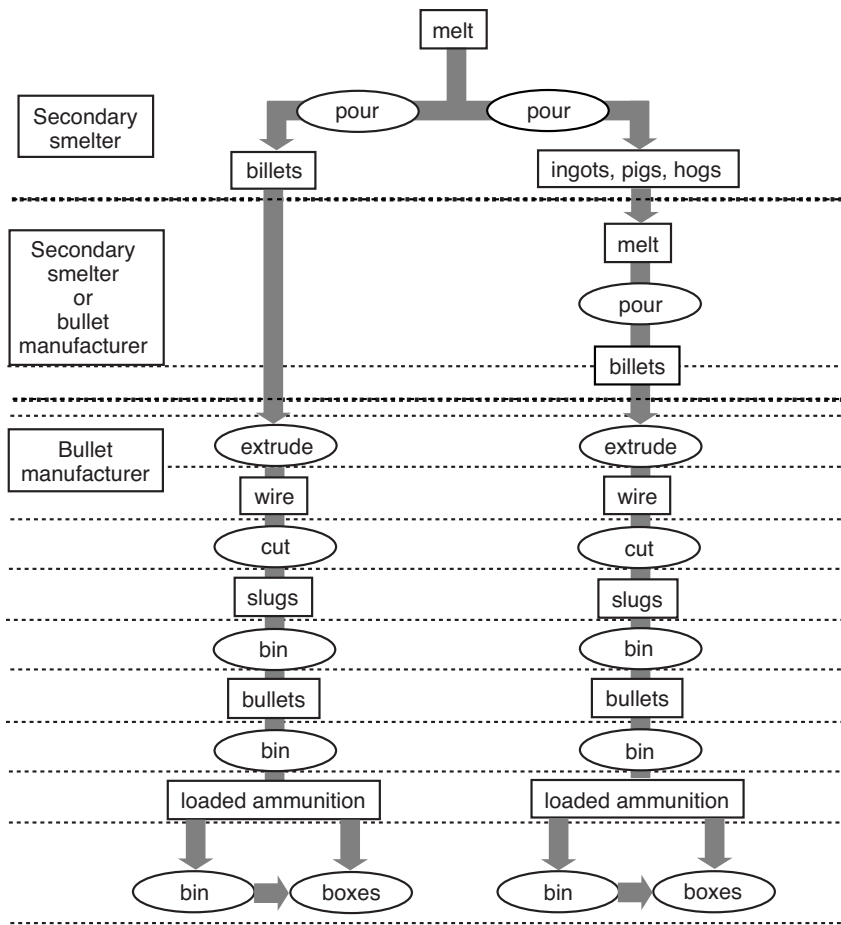
as many states; the courts have deemed that these criteria have been satisfied in the case of evidence based on DNA (which can be extracted from blood stains, human hair, etc.). But when the evidence consists of fragments of manufactured material such as glass, paint, or bullet lead, the Federal Rules of Evidence for admissibility are more controversial.

Bullet lead is one example of forensic evidence where the issues concerning admissibility raise much controversy. When a crime involving guns is committed and a gun is recovered, law enforcement officers try to match the striations on the gun barrel with those on the bullet casing; the validity and accuracy of methods for assessing a match are being studied currently. But when no gun is recovered, law enforcement officers send bullets found at the crime scene (CS bullets), along with bullets found in the possession of a potential suspect (PS bullets), to the crime laboratory at the FBI for compositional analysis of bullet lead (CABL). The "working hypothesis" justifying CABL is that the chemical concentrations of the lead used to make a 'batch' of bullets provide a unique



signature, so bullets that come from the same batch of lead should have the same concentrations of certain trace elements. One might be tempted to extend this unique signature concept to other manufactured material; accordingly, CABL raised several questions that would arise in connection with any crime scene evidence of this sort:

- (a) **Is the chemical profile of the batch really unique?** (If not, then statistical test sensitivity is compromised: the report can state no more than "match to any bullet produced from a batch with the same chemi-



**Figure 1. Flowchart of the bullet manufacturing process. Reprinted with permission from the NRC Report, p. 74.**

cal profile” i.e., the bullet matches any bullet from any one of potentially thousands of batches with the same chemical profile.

(b) **How large would a batch have to be before the batch size becomes so large the test is no longer specific to the batch from which bullets were made?** For example, few batches and millions of bullets produced from each batch would result in millions of bullets that could have matched the crime scene bullets, thus reducing the specificity of the test.

(c) **What is the appropriate statistical test for assessing a match?** This question requires consideration of measurement error, effects of the manufacturing process that produced the bullets (within-batch

and between-batch variability), and acceptable error rates of false positives (match when, in fact, the bullets came from different batches) and false negatives (no match when, in fact, the bullets came from the same batch).

(d) **How should “multiplicity” be taken into account?** That is, some cases produced only one or two CS bullets or bullet fragments and only a few PS bullets; other cases involved 10–20 CS bullets or bullet fragments and dozens of PS bullets. The number of possible pairwise comparisons may be small in one case, but large in another. How should we account fairly across all cases for multiple hypothesis tests?

(e) **Are the procedures for conducting CABL and the subsequent statistical analyses followed**

**consistently for all cases?** The error rates in statistical analyses are valid in the long run if the same procedures are used for each case.

These and many other issues raised concerns in the legal community regarding the scientific validity and legal admissibility of CABL in court proceedings. In 2003, the FBI commissioned a study of CABL, conducted under the auspices of the National Research Council, which assembled a panel of unbiased experts (seven analytical chemists, two forensic scientists, two forensic attorneys, one metallurgist, and two statisticians). The Committee on Scientific Assessment of Bullet Lead Elemental Compositional Comparison met four times in 2003 to address the FBI’s specific charge (NAS report, p.2, National Research Council (2004)):

- (1) Analytical method: Is the method (inductively coupled plasma optical (oratomic) emission spectroscopy—ICP-OES or ICP-AES) analytically sound and the best available? Is the selection of elements appropriate, and would useful information be gained by measuring isotopes?
- (2) Statistical procedures: Are the statistical tests used to compare two samples appropriate?
- (3) Borrowing strength: Can known variations introduced in the manufacturing process be used to model specimen groupings and provide improved comparison criteria?
- (4) Interpretation issues (probative value): What are the appropriate statements that can be made to assist the requester in interpreting the results of compositional bullet lead comparison? Can significance statements be modified to include effects of such factors as the analytical technique, manufacturing process, comparison criteria, specimen history, and legal requirements?

The process by which bullets are manufactured is described thoroughly in a report on this topic issued by the National Research Council in February 2004 (hereafter denoted as “NRC report”). Briefly, the process varies by manufacturer (the four largest bullet manufacturers in this country are Cascade Cartridge, Inc., or CCI; Federal;

Remington; and Winchester), but generally follows the flow chart shown in Figure 1. Most bullets are made from molten recycled lead (usually from car batteries); this lead is poured into smaller units—known as billets, ingots, or pigs—and then extruded into wires, cut into slugs of size appropriate to the caliber, and dumped into large bins that can hold tens of thousands of bullets. Compositional changes in the lead can occur during casting; for example, a manufacturer is likely to pour molten lead from other ingots into the pot to keep a continuous stream of bullet production. Robert D. Koons and Diana M. Grant noted one instance where the concentration of tin decreased 60% (from 300 parts per million to 120) over the course of 30 minutes. Consequently, the most homogeneous step of the process is the molten lead, which can yield as many as 35 million bullets; even when the unit of lead is smaller (billet), it often is large enough to produce 12,000 to 60,000 .22-caliber bullets. Bins usually are filled in the order in which bullets are produced, but, not uncommonly, they also can contain bullets made from different melts.

Bullet manufacturers are reluctant to share their data, so batch uniqueness could not be assured, leaving issue (a) open to debate. Also, the diagram of the flow chart of the manufacturing process (Fig. 1) confirms that a “batch” could contain thousands, or even millions, of bullets, so a bullet found in one box could match bullets from thousands of boxes. Thus the batch could be so large as to render virtually impossible the ability to identify a bullet with a specific box of 25-50 bullets (issue (b)).

Parts 2 and 3 of the charge most directly involve the validity of the application of statistical principles and methodology. Part 3, though not stated as such, seems to be directed at the evaluation of the FBI’s proposed method of clustering bullets of apparently similar chemical compositions—called “chaining”—designed to yield a larger sample of bullets with seemingly similar composition so means and standard deviations can be computed with a higher degree of confidence than those based on only the bullets at hand; in theory, such bullets could then be used as a ‘reference’ population for

purposes of comparisons with bullets of unknown sources.

Though the statistical principles for analyzing data from CABL are fairly straightforward (apart from choices of methods to estimate error variances, account for multiplicity, etc.), insufficient data were available to characterize within-batch and between-batch variability, which likely depended on manufacturer, caliber, origin, etc. The statistical analyses in the NRC report indicated within-batch variability appeared to be less than the FBI crime laboratory’s ability to measure it (measurement uncertainty) and often substantially less than the between-batch variability (though not always due to the “continuous stream of lead pours” in manufacturing bullets).

Statistical methods used by the FBI as part of CABL are discussed at greater length in the next section; the adoption of appropriate statistical tests requires consideration of acceptable false positive and false negative rates, decisions that are notoriously difficult to resolve in legal contexts. Question (d), concerning multiplicity, can affect false positive and false negative rates, but it appears not to have been considered at all by the FBI.

Finally, (e) addresses the probative value of CABL, specifically whether CABL satisfies the criteria that govern the admissibility of evidence in federal courtrooms as well as my states, known as “Federal Rules of Evidence 702” (FRE 702):

If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

The NRC report further explains that the trial court must make a “preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether

that reasoning or methodology can be applied properly to the facts in issue.” In performing this gatekeeping function, the trial court may consider a number of factors: whether the theory or technique can be and has been tested, whether it has been subjected to peer review and publication, a technique’s known potential error rate, the existence and maintenance of standards controlling the technique’s operation, and a technique’s general acceptance in the relevant scientific community. Those factors, however, are neither dispositive nor exhaustive. The court emphasized that FRE 702 is “a flexible one.” The NRC report also quotes 509 U.S. (1993) at 593–594:

Publication (which is but one element of peer review) is not a *sine qua non* of admissibility; it does not necessarily correlate with reliability. ... The fact of publication (or lack thereof) in a peer-reviewed journal thus will be a relevant, though not dispositive, consideration in assessing the scientific validity of a particular technique or methodology. ... Widespread acceptance can be an important factor in ruling particular evidence admissible, and “a known technique which has been able to attract only minimal support within the community,” ... may properly be viewed with skepticism.

The three specific criteria identified in FRE 702—namely, (1) based on specific facts or data, (2) based on reliable principles and methods, and (3) consistent and reliable application of principles and methods—are discussed below with reference to the chemical and statistical analyses conducted on bullet lead. We argue that the background for these criteria—data and specification of methods and protocol—should be subjected to the same demanding standards as for any scientific method, namely review, confirmation, validation, and open access to data and methods. After describing the methods used in CABL, we discuss whether CABL satisfies the first criterion, “sufficient facts or data,” specifically with regards to (a) and (b).

The next section discusses whether the FBI’s statistical analyses used in

CABL satisfactorily address criterion (2), “reliable principles and methods,” specifically with reference to issues (c) and (d)—statistical tests and multiplicity. We then consider the third criterion of FRE 702, reliable (and consistent) application of principles and methods to the facts of the various cases (issue (e), consistency of analytical—chemical and statistical—procedures and potential for widely varying numbers of bullets and bullet fragments in each case). In each section, we also discuss other facts the court is permitted to consider as part of the “flexibility” in applying FRE 702 (e.g., peer review, error rate, existence and maintenance of standards, general acceptance). We conclude with an update on the current state of CABL in this country, the impact of scientific inquiry on other forms of evidence besides bullet lead, endorsements for the application of the scientific method in all areas of forensic science, and some lessons learned for statisticians who are asked to participate on future panels of this sort.

### CABL Procedures—Chemical and Statistical

When the local police department would send evidence to the FBI crime laboratory, the FBI examiner in charge would photograph and catalog the contents before conducting formal chemical analyses of trace elements. Prior to 1990, measurements were made by neutron activation analysis (NAA), a gamma-ray counting method using radioactive samples that decay over time. Later, elemental concentrations were determined by three consecutive scans of an optical spectrometer over wavelengths characteristic of the trace element in question (i.e., via ICP-OES or ICP-AES). Before 1990, usually four trace elements were measured: antimony, Sb; copper, Cu; silver, Ag; and bismuth, Bi. Three additional elements have since been added: arsenic, As; tin, Sn; and cadmium, Cd. The NRC committee determined that “the current analytical instrumentation used by the FBI is appropriate and is the best available technology with respect to both precision and accuracy for the elements analyzed in a lead matrix,” but also that “the current FBI procedure is not documented in a complete and detailed format that would allow other

laboratories skilled in the art to practice or even fully evaluate it.”

Although the details of the chemical protocol for obtaining the measurements were deemed by the NRC committee to be incomplete, the general formulation can be described here. To assess a match (analytically indistinguishable chemical compositions) between a CS bullet and a PS bullet, the lab technician would extract three pieces from each PS/CS bullet (or bullet fragment). Nominally, each of the three pieces was measured in triplicate, using three standards provided by the National Institute of Standards and Technology (NIST SRM 2415, 2416, or 2417), by ICP-OES. In practice, some cases included bullet fragments that were too small to yield three pieces, which resulted in only two measurements (still measured in triplicate). The three replicates on each piece were averaged, and then the sample means—standard deviations (SDs)—and ranges (minimum, maximum) of these (nominally three) averages were calculated for each of the seven elements for each CS and PS bullet. The committee had no triplicate measurements, so it had to accept the FBI’s claim that the error in replicates (using three standards) was substantially lower than that due to other sources (e.g., piece-to-piece variability within the same bullet). The averages (from the triplicates) on the (nominally) three pieces are called “measurements.” The report noted lack of detail in the FBI Crime Lab’s protocol for measuring the concentrations.

Regarding the statistical procedures, suppose that  $\sigma_e$  represents the true measurement (analytical) error. (Chemists recognize that “relative error” is a more useful concept for their measurements, so actually  $\sigma_e$  will represent an average standard deviation of the measurements on the log scale.) The FBI protocol for assessing a match between these two bullets specified a “2-SD procedure”: if the “2-SD-interval,” namely  $(\text{mean} \pm 2 \cdot \text{SD})$  calculated on the PS bullet, overlaps with the 2-SD-interval for the CS bullet, then a match is declared. In the best case scenario, the (log) measurements are perfectly Gaussian with no outliers, and  $E(\text{SD}) = 0.8862\sigma_e$  (based on  $\chi_2^2$ ), so this procedure declares match when the means are, on average, within  $3.5448\sigma_e$ . Comparing this quantity with that obtained using a con-

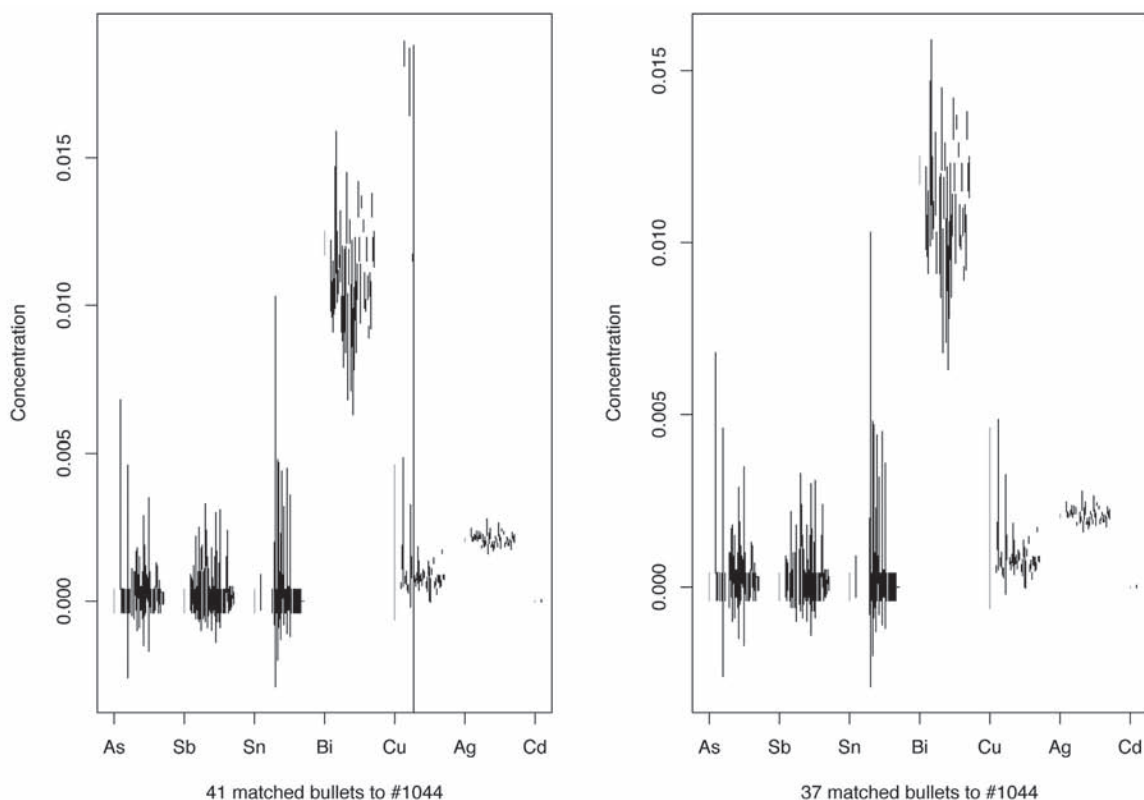
ventional t-test based on pooled standard deviations (the FBI had conducted CABL on many thousands of bullets, so degrees of freedom in such pooled SDs would be very large). An interval of width  $3.5448\sigma_e = z_{\alpha/2}\sigma_e\sqrt{1/3+1/3}$  corresponds to significance level  $\alpha = 0.000014$ . Adjusting for multiplicity might make this level appropriate, except for the facts that the seven tests on the seven elements were not independent (the estimated correlation matrix was far from the identity) and the null hypothesis in this case should probably be “means are not equal” (i.e., suspect is innocent), not the conventional “means are equal” (i.e., suspect is guilty). (The fact that the FBI calculated sample means and SDs on the raw scale likely rendered the intervals even wider, as the SD would be inflated. Usually, the chemists report measurement uncertainty as “relative SD,” but the FBI reported the SD on the raw measurements for the 2-SD-overlap match procedure.)

Although formal reports on bullet comparisons submitted by the FBI appeared to be based on only the results from the 2-SD overlap procedure, the FBI also occasionally used a “range test,” which declared a match between two bullets if their ranges overlapped. Because the expected range of three Gaussian (log) measurements is  $0.8463\sigma_e$ , this procedure declares a match whenever the difference in the sample means lie, on average, within  $1.6926\sigma_e$  of each other. Not surprisingly, the FBI noticed fewer matches with the “range overlap” test than with the 2-SD overlap test.

The FBI also used a chaining procedure to construct ‘groups’ of ‘similar’ bullets: If the 2-SD intervals on all seven elements for one bullet overlapped with all seven 2-SD intervals for another bullet, then the two bullets were declared to lie within the same ‘group.’ Clearly, chaining can result in grouping some bullets whose intervals on all seven elements are completely disjointed. This effect is demonstrated in Figure 2.

### Sufficient Facts or Data

A proper evaluation of the FBI’s statistical procedures, such as their 2-SD overlap method for identifying matches, required some assessment of the consistency of the (log) chemical measure-



**Figure 2.** Illustration of the consequences of “chaining.” Panel (a) shows 2-SD interval for bullet #1044 (from the 1,837-bullet dataset) as the first line in each set of elements, followed by the 2-SD interval for each of 41 bullets whose 2-SD intervals overlap with that of bullet #1044. Four of these 41 bullets had extremely wide intervals for Cu, so they are eliminated in Panel (b). Another 2-SD interval was constructed from the SD of the 42 (38) bullet averages on each element, resulting in a total of 58 (57) bullets that matched. Reprinted with permission from the NRC Report, p.203.

ments to follow a Gaussian distribution with a consistent  $\sigma_e$ . The FBI provided three datasets to assist the NRC committee with these tasks.

#### 800-Bullet Dataset

This dataset arose from a study to characterize the variability in the number of “analytically distinguishable groups” in a typical box of bullets. The study included measurements of concentrations of 4, 5, or 6 trace elements, using NAA and/or ICP-OES, on all 50 bullets in each of four boxes from the four major U.S. bullet manufacturers—a total of 800 bullets. Six of the seven trace elements (all but Cd) were measured by ICP-OES for only the 200 Federal bullets; all other bullets were measured for fewer than six elements, some by NAA and some by ICP-OES. The bullets did not come from any real case; thus, the resulting estimates of the measurement of uncertainty are likely to be somewhat

optimistic due to the potential for bullets that are part of a research study to be measured somewhat more carefully. This was the only dataset received by the committee in a timely fashion that contained the three individual ICP-OES measurements on six of the elements (all but Cadmium) and therefore the only dataset from which correlations among measurement errors of the six elements could be estimated. Nearly all the other datasets provided to the committee reported only means and SDs on the three measurements. Hence, only somewhat ‘optimistic’ estimates of  $\sigma_e$  and pairwise correlations were available at the time the NRC report was published.

#### 1,837-Bullet Dataset

The FBI created this dataset to estimate the false match probability (false positive probability) in a typical FBI CABL case. An honest evaluation of

this probability would require a random sample of all manufactured bullets that were known to have been manufactured from separate lots. Rather than obtaining such a sample, however, the bullets in this data file were selected to include one (or sometimes more) bullets that were “assessed” to be distinct from the other bullets in the case; a few are research samples “not associated with any particular case” and a few “were taken from the ammunition collection” (again, not associated with a particular case). Quoting from the notes that accompanied this dataset:

To assure independence of samples, the number of samples in the full database was reduced by removing multiple bullets from a given known source in each case. To do this, evidentiary submissions were considered one case at a time. For each case, one specimen from each combination of

bullet caliber, style, and nominal alloy class was selected and that data was placed into the test sample set. In instances where two or more bullets in a case had the same nominal alloy class, one sample was randomly selected from those containing the maximum number of elements measured. The test set in this study, therefore, should represent an unbiased sample in the sense that each known production source of lead is represented by only one randomly selected specimen.

Clearly, this dataset does not represent a random sample of bullets known to be different. (In fact, the application of chaining to bullet #1044, shown in Figure 2, resulted in 44 matches, demonstrating both that chaining can group apparently different bullets and that the dataset may have included bullets from the same source.) Evidence for lack of “randomness” comes from Table 3 of Koons and Basaglia, who report a total of 3 of the 1,837 bullets that come from CCI (one of the largest U.S. bullet manufacturers), 3 from Federal, and 361 bullets (20%) from China or Korea.

Due to the lack of a consistent dataset for assessing the probability of a false match, the committee resorted to simulation for quantifying this probability. Not surprisingly, the simulation demonstrated, under utopian conditions (Gaussian distributions and constant measurement error), that the false positive rate can be as high as 4% in the utopian case of independent measurement errors and 9% under the more realistic case of mildly correlated errors when the true difference exceeded the measurement error by a factor of 3. (Smaller differences result in higher false positive rates, while larger differences yield lower false positive rates.) This dataset was useful only for providing some sense of typical trace element concentrations, but not for assessing definitive characteristics about distributions or correlations between measured concentrations or for assessing error rates of any statistical procedure.

#### *‘Complete’ Data File*

During the open sessions of the committee meetings, the FBI claimed to have a “complete data file” of some 71,000+ measurements. Following repeated

requests from the committee, the FBI submitted at its last meeting a CD-ROM that contained two data files with a combined total of 64,869 bullet (not 71,000+) measurement records. This dataset could not be analyzed in time for the release of the report; however, subsequent inspection of it identified several peculiarities. For example, the dataset contained only measurements made via ICP-OES (recall, NAA was used prior to 1997). Second, the numbering system of the bullets was highly inconsistent and rather unexpected (e.g., the bullets from a suspect in a particular case might be numbered Q13A, Q13B, Q13C, Q14A, Q14B, Q14C, ..., leading one to wonder what happened to bullets Q01, Q02, ..., Q12). Third, while most of the bullets indicated three measurements, about 30 bullets had six or more measurements. Fourth, a rough investigation of the measurement error indicated many measurement errors that exceeded the FBI’s claimed analytical precision of 2–5%. Fifth, only about 50% of the bullets in this dataset were identified as having come from one of the four major bullet manufacturers in the United States; the “complete data file” of 71,000 bullets may yield a higher proportion of bullets from these four manufacturers. Sixth, only 15% of the 1,079 cases listed in these two files had measurements from NIST standards (S), PS bullets (Q), and CS bullets (K); most cases had only S and Q bullets—or only S and K bullets—making it impossible to determine the frequency of matches between PS and CS bullets in a case. The missing data and the inconsistent precisions lead one to question the validity of the “sufficient facts or data.”

### **Reliable Principles and Methods**

As indicated above, the FBI’s “2-SD procedure” for assessing a match between two bullets was not consistent with proper and validated statistical procedures. When the results of CABL are presented to the jury, the jury must decide whether—given the data—the defendant is more likely to be guilty than innocent. Conventional hypothesis tests would dictate the use of equivalence tests for this situation: are the results from CABL “too close” for one to believe that the bullets really came from

separate sources? Equivalence tests do require some specification or “limit” of how different the bullets can be and still be considered to have been from the same source (i.e., if  $\delta$  represents the true difference in the bullet means, then the null hypothesis would be stated as  $H_0: \delta > \delta_0$ ;  $\delta_0$ , must be prespecified, and usually is based on the analytical precision of the measurement technology).

The FBI’s “2-SD-overlap” test corresponded roughly to a series of seven (somewhat correlated) equivalence  $t$  tests (or a multivariate equivalence  $T^2$  test) with an overall significance level of 0.001 (i.e., one false match, on average, for every 1,000 pairs of CS/PS bullets tested) if  $\delta_0 = 3.3\sigma_e$  (i.e., “same batch” meant “within 3.3 times the measurement uncertainty”). Given the expected homogeneity of the material within a batch and the crime lab’s sometimes large measurement uncertainty, this constitutes a rather generous definition of “equivalent batches,” resulting in quite a few claimed matches.

A likelihood approach is no more helpful in this context; it would require the jury to determine whether the ratio

$$\frac{P\{\delta < \delta_0 \mid \text{statistical test claims “match”}\}}{P\{\delta > \delta_0 \mid \text{statistical test claims “match”}\}}$$

exceeds one (i.e., How likely are the bullets to have come from the same source, given the results of the statistical test?). Unfortunately, Bayes rule shows the probabilities in this ratio depend on the distribution of  $\delta$  in the population of bullets. In fact, the distribution of  $\delta$ ’s in the population likely depends upon caliber, manufacturer, geographical location, and a host of other factors. None of the available datasets provided reliable, unbiased information to accurately quantify this difference.

### **Reliable and Consistent Application of Methods**

To assess whether the statistical methods were applied consistently and reliably to various cases, the committee also requested from the FBI a list of those cases where a match was found and which CS and PS bullets matched. The 64,869-record file did not indicate matches and nonmatches. In theory, the statisticians on the committee might have attempted to reconstruct the FBI’s

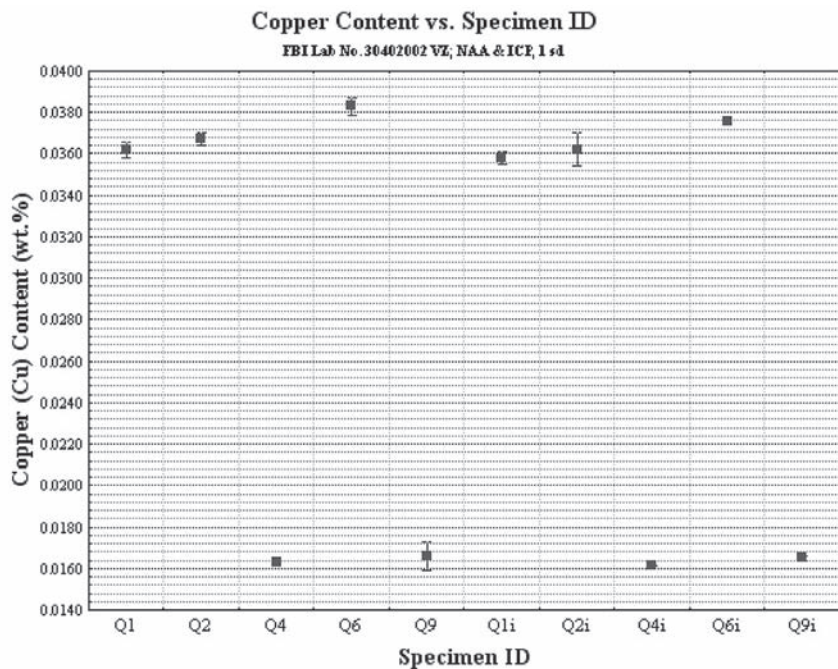
“2-SD-interval” procedures on every pair of CS and PS bullets within a case, but would not have been able to confirm or deny any matches they had found with those found by the FBI. (The FBI did agree to pull the records for any case about which the committee had questions.) However, an example of inconsistency in testimony appears in Figures 3 and 4, the measured concentrations of copper (Figure 3) and arsenic (Figure 4) from a case in which the defendant was found guilty based in part on CABL evidence. In these figures, Q1 is a CS bullet (extracted from the victim), Q2 and Q6 are PS bullets (found in the suspect’s possession), and Q4 and Q9 are not specified as either CS or PS bullets. (The “i” following the bullet label—viz., Q2i—indicates measurement via ICP-OES, instead of NAA.) One-SD intervals are shown, but it is clear that even 2-SD intervals surrounding Q1 (CS) and Q6 (PS) would not overlap for Cu, and Q1 and Q6—as well as Q1 and Q6—would not overlap for either Cu or As. Nonetheless, the testimony of C. Peele in Baltimore County, Maryland, on November 16, 1995, leaves the impression that Q1 matches bullets found in the suspect’s possession:

**Q.** And, and going down to Q5 and Q8, again, those numbers are quite varied from Q1 and Q2, correct?

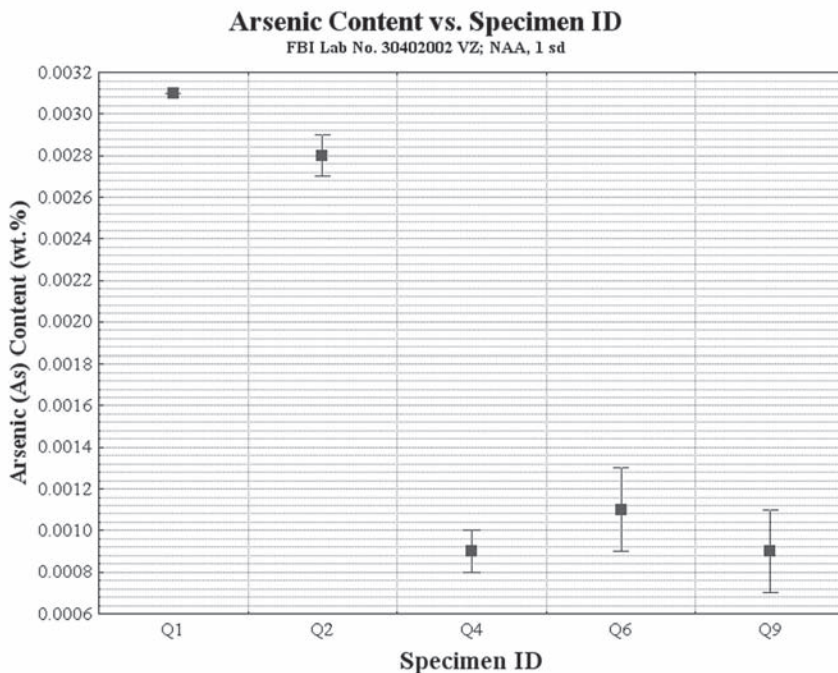
**A.** No, not quite varied. Those are varied. In other words, they are different in composition. But all these differences are not very large. As you can see, some are quite a bit less than others. The difference between the five, eight, and all the remaining ones are easily measurable. But, certainly, those are differences that can be expected. Even in one box of ammunition, not every bullet in a box is the same. As a matter of fact, there are usually a number of different compositions in one box. And this is what you can see even in one box.

**Q.** However, you cannot state that samples Q4 through Q9 are consistent with Q1 and Q2 as having come from the same box, correct?

**A.** Q6 is much more so than any of the rest. Q6 is so close that, certainly, that could have been in the same box.



**Figure 3.** Copper concentrations in bullets from an actual case. Q1 is a bullet from the crime scene; Q2 and Q6 were found on the suspect’s property; Q4 and Q9 are unknown. The “i” following the bullet label indicates measurement via ICP-OES instead of via NAA.



**Figure 4.** Arsenic concentrations in bullets from an actual case. Q1 is a bullet from the crime scene; Q2 and Q6 were found on the suspect’s property; Q4 and Q9 are unknown. The “i” following the bullet label indicates measurement via ICP-OES instead of via NAA.

The NRC report quotes from the testimony in several cases over the years, which leaves the impression that the results of CABL were more definitive than could be justified in practice (e.g., “could have come from the same box”; “must have come from the same box or another box that would have been made by the same company on the same day”; “had come from the same batch of ammunition: they had been made by the same manufacturer on the same day and at the same hour”). Based on these examples, one would have to question whether CABL satisfies “reliable and consistent application of methods to facts of various cases.”

## Postscript

The main thesis of this paper is that forensic scientists and examiners have the same responsibility as other scientists when results are presented in public forums. Those forums may include archival journals, public meetings, news conferences, courts proceedings, or any other venue where scientific findings are given to the public. Material, methods, and data used to obtain and support a scientific claim should be made available to responsible parties in an appropriate manner.

The executive summary of “Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences” emphasizes the importance data sharing:

Community standards for sharing publication-related data and materials should flow from the general principle that the publication of scientific information is intended to move science forward. More specifically, the act of publishing is a quid pro quo in which authors receive credit and acknowledgment in exchange for disclosure of their scientific findings. An author’s obligation is not only to release data and materials to enable others to verify or replicate published findings (as journals already implicitly or explicitly require) but also to provide them in a form on which other scientists can build with further research. All members of the scientific community—whether working in academia, government, or a com-

mercial enterprise—have equal responsibility for upholding community standards as participants in the publication system, and all should be equally able to derive benefits from it.


Certainly, this statement applies to CABL as well as to all aspects of forensic science.

The scientific method is important for science generally; forensic science is no exception. Using CABL as one example of an area of forensic science, the evidence in this paper suggests that, at least for CABL, forensic science failed in the requirement to share the materials, methods, and data used to reach conclusions with the scientific community. How can compliance with these standards be encouraged? Rather than discrediting CABL directly, the final report of the NRC Committee noted the lack of CABL’s adherence to the scientific method, the gaps in knowledge and data, and the failure of CABL to meet the conditions specified by the Federal Rules of Evidence. The NRC held a press conference announcing release of the report and various panel members accepted invitations to discuss the report in public forums (e.g., Joint Statistical Meetings in 2004; Spring Research Conference in 2005) and publish articles in peer-reviewed journals. How successful were these efforts in encouraging the FBI to reconsider the use of CABL as forensic evidence?

Shortly before the committee’s report was released, the legal counsel for a defendant standing trial in U.S. District Court, N.D. Illinois Eastern Division, argued successfully to preclude the introduction of CABL evidence into testimony on the basis that CABL failed to meet satisfactorily Federal Rules of Evidence 702. Although some of the statements in this motion are statistically misguided (“a sample of 1,837 is extremely small to reliably extrapolate principles as to the total bullet population”), the motion did call attention correctly to the lack of reliable, consistent application of CABL and inferences from it.

The NRC report was released in February 2004. In March 2005, a New Jersey appeals court overturned a 1997 murder conviction, stating that the NRC report “raised new questions about the technique the FBI has used for decades

to match bullets to crimes.” Six months later, the FBI announced it had decided to abandon its use of CABL. According to an Associated Press article on September 1, 2005, “The FBI said its decision to drop the tests was significantly influenced by the fact that ‘neither scientists nor bullet manufacturers are able to definitively attest to the significance of an association made between bullets in the course of a bullet lead examination.’” By insisting on data sharing, integrity, and proper evaluation of forensic procedures, the scientific community successfully influenced the FBI to abandon its use of CABL as forensic evidence in the courtroom.

Formal laws demanding adherence to the scientific method may or may not be successful. The story of CABL may suggest instead a more successful route to encourage compliance: assemble panels of experts to evaluate these methods, bring the available data and methods to the attention of the scientific community, and discuss the results in peer-reviewed publications and public forums. Other sorts of evidence have been used in the same vein as bullet lead. We hope the lessons learned from CABL about the importance of using the scientific method to its analysis will be applied for those situations also. 

## References

- Geisser, S. (2000) “Statistics, Litigation, and Conduct Unbecoming,” in *Statistical Science in the Courtroom*. New York: Springer.
- Grant, D.M. (2003) Personal communication to committee, April 2003.
- Koons, R.D. (2003) Notes on the 1,837-bullet database. Committee communication.
- Koons, R.D. and Basaglia, J. (2005) “Forensic Significance of Bullet Lead Compositions,” *Journal of Forensic Science*. 50(2).
- Koons, R.D. and Grant, D.M. (2002) “Compositional Variation in Bullet Lead Manufacture,” *Journal of Forensic Science*. 47:950–958.
- National Research Council (2003) *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Washington, DC: The National Academies Press.

National Research Council (2004) *Forensic Analysis: Weighing Bullet Lead Evidence*. Washington, DC: The National Academies Press.

Peele, E.R., Havekost, D.G., Peters, C.A., Riley, J.P., Halberstam, R.C., and Koons, R.D. (1991) "Comparison of Bullets Using the Elemental Composition of the Lead Compo-

nent," *Proceedings of the International Symposium on the Forensic Aspects of Trace Evidence*.

Peters, C.A. *Comparative Elemental Analysis of Firearms Projectile Lead by ICP-OES, FBI Laboratory Chemistry Unit*. Unpublished.

Randich, E., Duerfeldt, W., McLendon, W., and Tobin, W. (2002),

*Forensic Science International*. 127:174–191.

Tobin, W. (2005) Personal communication.

Wellek, S. (2003) *Testing Statistical Hypotheses of Equivalence*. New York: Chapman and Hall.

---

## Comment: Further Arguments against CABL as a Forensic Tool

---

Alicia Carriquiry, Iowa State University

**Editor's Note:** *Collaborators in the earlier work described here include Hal Stern, Michael Daniels, and several graduate students.*

Is it really any wonder the FBI has finally given up on CABL as a means to cast doubt on a suspect's innocence? The method of CABL, extraordinarily well described in "Data Integrity and the Scientific Method: the Case of Bullet Lead Data as Forensic Evidence," has been under fire for many years. Arguments against the reliability of the method and—more importantly—against the sweeping conclusions that have been presented as truth in courts of law have finally had an effect. Spiegelman and Kafadar nicely summarize some of the controversial issues surrounding the implementation of CABL as a tool to assess the guilt or innocence of a suspect. I very much enjoyed their careful discussion of the shortcomings in the statistical methods employed by FBI forensic scientists to 'match' two or more bullet lead specimens recovered from a crime scene and from a suspect. In this discussion, I offer a few additional comments about the specific case of CABL as a forensic tool and leave for the end commentary about what is perhaps the most important point in the Spiegelman and Kafadar article: that only open scientific and public scrutiny can guarantee the integrity of the data and the methods used to draw inferences about any problem in just about any venue, and that some current practices in expert testimony go contrary to those principles.

When a crime is committed and evidence is recovered from the crime scene and from a suspect, assessing the degree of culpability of the suspect hinges on the answers to two questions: Can we say with some degree of certainty that the samples recovered at the crime scene and on the suspect are a match? If samples match, could the match have occurred by chance alone? In other words, is it possible, or even likely, that someone other than the suspect could have deposited the evidence at the crime scene?

If we can confidently answer no to the second question, we would conclude the evidence has high probative value, as a match would suggest strongly that it was the suspect (and almost no one else) who could have deposited the sample at the crime scene. Arguably, none of the two questions can be answered with enough confidence in the case of bullet lead, thus casting enough doubt on CABL as a forensic tool as to render it useless. In all fairness to the FBI, doubts about the scientific and statistical underpinnings of CABL as forensic evidence arose as early as 1998. At that time, the FBI funded a study carried out at Iowa State University, which resulted in a report that became distributed rather widely. In that report, Michael Daniels, Hal Stern, and I (2000) suggested the two questions above could be answered simultaneously by evaluating a posterior odds ratio of the form

$$\frac{\Pr(G|E)}{\Pr(\bar{G}|E)} = \frac{\Pr(E|G) \Pr(G)}{\Pr(E|\bar{G}) \Pr(\bar{G})}, \quad (1)$$

where (simplistically)  $G$  denotes 'guilt,'  $\bar{G}$  denotes its complement (or 'not guilt'), and  $E$  stands for evidence. The ratio on the left side of the equation estimates the relative probabilities of guilt and not guilt, given the evidence. The first ratio on the right side, equivalent to a likelihood ratio for the data or evidence,

quantifies the relative probabilities of observing the evidence under the two competing hypotheses: that the suspect is guilty or that she isn't. It is this likelihood ratio that we wish to evaluate to assess the probative value of evidence  $E$ . While the actual form of the numerator and denominator of the likelihood ratio will depend on the specific application, the probative value of any type of evidence can, in principle, be assessed by estimating the ratio from actual data and from our knowledge about the processes followed to collect and analyze those data. Alas, estimating the likelihood ratio in all cases except the trivial ones is not an easy task.

Let us first consider the issue of the match, closely associated to the problem of estimating the numerator of the likelihood ratio in (1). Recall that the numerator in the ratio estimates the probability we have observed the evidence given that the suspect has deposited the evidence at the crime scene. That the numerator in the LR ratio measures the probability of a match is easy to see if we consider, for example, blood type evidence. The numerator in the LR would then quantify the probability of observing the suspect's blood type on a biological sample recovered from the crime scene, given that the suspect left that biological sample. Under the hypothesis of guilt, we expect the blood type from the crime scene sample to be indistinguishable from the suspect's blood type. Spiegelman and Kafadar, and to a large extent the NRC report, largely focus on this issue (although they also touch upon the issue of a coincidental match).

Because it often is not possible to measure anything without some measurement error, and because variability across seemingly 'identical' units is almost inevitable, statistics must play a fundamental role in this first step. While it might be possible to disagree on the

actual *criteria* that can be invoked to determine whether two samples match, the *statistical methodology* implemented to decide whether those criteria are met is either correct or it is not. Spiegelman and Kafadar do a wonderful job of dissecting the many questionable statistical practices employed by FBI analysts to establish matches and the difficulties that would arise if one wished to actually carry out a correct statistical analysis of data arising from CABL studies. It almost goes without saying that implicitly assuming normality of severely skewed measurements, ignoring correlations among multiple measurements collected on the same unit and implementing ad-hoc methods—such as chaining—to identify groups of ‘indistinguishable’ samples, can lead to misleading conclusions and is clearly in violation of the Federal Rules of Evidence.

The real limitations of CABL, however, become glaring when we attempt to estimate the denominator of the LR in (1) to provide an answer to the second question. In statistical terms, the problem consists in estimating the probability that the match between the crime scene and the suspect’s evidence has been established, *even though the suspect was not in contact with the crime scene*. In other words, the denominator in the LR quantifies the probability of observing a match under the hypothesis of no guilt.

Consider again our blood type example and suppose a blood sample found at a crime scene in Ames, Iowa (population of about 50,000), is found to be of type A+. A suspect, who also has blood type A+ is tried for the crime. In the absence of other incriminating evidence, the suspect is almost certain to be set free, as he and approximately 17,000 other Ames residents (or, on average, 34% of the population) might have deposited that blood at the scene. Blood types have low probative value because they are not *rare* and thus result in a large value of the denominator of the LR in (1). We know this is the case because *population studies* have been conducted and have allowed estimation of the approximate proportions of individuals who can be expected to carry each blood type. What is the value of the corresponding denominator in the case of CABL? Your guess is as good as anyone else’s, as no population studies have ever been conducted on bullets. Thus, no reference population

parameters against which to compare the sample characteristics have even been estimated.

Surprisingly, this issue has received relatively little attention in much of the discussion surrounding the validity of CABL. The NRC report mentions this issue almost exclusively in connection to the number of bullets that might be manufactured from a single, more or less homogeneous batch of lead and to the possibility that similar trace element concentrations may be repeatable and occur across various batches of lead. But the probability of a coincidental match in bullet lead is driven by many other factors as well, and these other factors may in fact be the proverbial elephant in the room. One such factor is the geographic distribution of bullets, a notoriously difficult one to tackle and that weighs directly on the relevance of whatever reference population of bullets is used to estimate the denominator in (1). Given that crimes are committed in localities with different numbers and types of ammunition retailers, who sell vastly different numbers of bullets in any period of time and receive their shipments from a varying assortment of manufacturers with distribution warehouses located in different places, it seems reasonable to conclude a different reference population would need to be established for almost each and every crime, as variation in the ‘local population’ of bullets from which the suspect could have purchased hers is likely to occur between town and town and from one month to the next. Two bullets bought by two individuals might be more likely to match if both were purchased in the same month at the only retailer in a tiny town than if both were purchased in a large city with a large number of retailers with high volumes of sales. As a consequence, the probative value of CABL evidence might well be different for the small town and the big-city suspects.

The validity of CABL as forensic evidence has been dealt a serious blow, and this is a good thing. What is not so good, however, is that other types of evidence routinely presented at trial deserve, but have not received, the same type of scrutiny. Most of the concerns raised about CABL can be raised as well (with other names and slightly different emphases) about glass fragments, paint chips, and carpet fibers. In contrast to

‘biological evidence,’ for which population parameters can be expected to be more or less stable over time (perhaps within the appropriate subpopulations), manufactured products can be expected to change over time and across brands and locations. Thus, reference populations needed to estimate probabilities of coincidental matches are difficult to define, sample from, and interpret.

One final and important point in the Spiegelman and Kafadar article deserves attention. The authors argue that abiding by the principles of the scientific method requires that access to the data and the technology employed to evaluate forensic evidence be made available to the scientific community. This, unfortunately, is not likely to happen any time soon. Much of today’s forensic analyses require the application of sophisticated analytic and quantitative methodologies. Often, these methods are implemented using patented, licensed, or otherwise restricted technologies that are sometimes (understandably) jealously protected by industries or groups holding those patents and licenses. Consider, for example, ballistics. One goal in ballistics is to decide whether two or more bullets could have been fired by the same gun by comparing the markings left on the casing by the gun’s firing pin. To carry out ballistic analysis, an image is obtained first from all samples to be compared, and then the images are compared using quantitative methods. ATF (Bureau of Alcohol, Tobacco, and Firearms) firearms examiners, as well as local law enforcement, use an automated system that helps them carry out those comparisons in a much more efficient manner. There is one vendor for the software used in ballistic evaluations in the United States. The software is licensed to law enforcement as a ‘black box,’ in the sense that details about the algorithms implemented by the program are kept largely a secret.

Even if proprietary technology is trustworthy and has been shown to produce reliable results, the lack of public and scientific scrutiny of this type of technology can be problematic. If nothing else, participation of a wider segment of the scientific community in the development of improved versions of any technology can accelerate the pace at which progress is made and can be nothing but good for society at large. 