

# Evaluating and Challenging Forensic Identification Evidence



Television dramas like the popular “CSI” series have highlighted the importance of forensic science in criminal investigations. These programs show forensic scientists solving crimes with unerring accuracy by examining and comparing bloodstains, hairs, bullets, glass fragments, handwriting, toolmarks, latent prints and other items of physical evidence. Ironically, while television has been glorifying crime labs, there has been growing skepticism among real scientists about the claims that forensic scientists have been making in court.<sup>1</sup> The conclusions they state so confidently on television and in real courtrooms often cannot withstand critical scrutiny. A recent article in the prominent journal *Science* argued that the “forensic identification sciences” are largely “underresearched and oversold” and called for major new efforts to test and validate claims that forensic scientists have routinely been making in courtroom testimony.<sup>2</sup>

For criminal defense lawyers, these developments are both a challenge and an opportunity. It will not be easy to dispute the prevailing wisdom, fed by CSI-style media fantasies, that forensic science is virtually infallible. Yet the intellectual weaknesses of many of the “forensic sciences” are now becoming increasingly apparent. Accordingly, it is timely to take a fresh and skeptical look at forensic evidence of all types. This article will present a general framework for evaluating the often nebulous and questionable claims of forensic experts. It will dis-

cuss strategies and techniques that can be used to evaluate any type of forensic identification evidence.

In developing the general framework, we will draw many lessons from the recent successful challenge to Comparative Bullet Lead Analysis (CBLA), a forensic technique that the FBI used in criminal investigations since the 1960s to link bullets found at crime scenes to

boxes of ammunition owned by suspects. FBI analysts used sophisticated instruments to measure the level of various elements in the lead alloy of each bullet in order to develop a chemical profile of the bullet. If the profile or the crime scene bullet “matched” the profiles of bullets in a box of ammunition owned by a suspect, the expert would conclude that the crime scene bullet came from the same molten source at the manufacturer and, frequently, from the same box. CBLA was particularly important in cases where the crime scene bullet was too damaged or fragmented to compare to a particular gun, or where no gun was recovered.

Although FBI analysts have been testifying about bullet lead “matches” since the 1960s, the validity and probative value of CBLA evidence was only recently called into question.<sup>3</sup> As a result of the challenge, on September 1, 2005 the FBI announced that it would no longer continue this type of testing.<sup>4</sup> Because the FBI operated the only laboratory in the United States that routinely performed CBLA, its decision

By William A. Tobin And William C. Thompson

effectively ends the technique, at least for now. According to FBI Laboratory Director Dwight Adams, the decision to discontinue CBLA was “based primarily on the inability of scientists or manufacturers to definitively evaluate the significance of an association between bullets made in the course of a bullet lead examination.”<sup>5</sup> In other words, the FBI decided there is no point in continuing to generate CBLA evidence because no one can say for sure what this evidence means. This development raises two questions that we invite readers to contemplate. Why did it take so many years to recognize this fatal weakness in bullet lead evidence? And how many other types of forensic evidence might have similar flaws or weaknesses?

In developing our general framework for analyzing forensic evidence we will also draw lessons from DNA evidence, which has been more thoroughly and carefully validated than most other areas of forensic identification science. Many fundamental questions about the meaning and value of DNA have been addressed with solid scientific research while similar questions about other types of forensic evidence are left to intuition and guesswork. While DNA evidence is hardly infallible, and the results of DNA tests must be carefully scrutinized to assure the tests were conducted and interpreted properly, it nevertheless serves as a useful point of comparison with the less advanced areas of forensic identification science.

#### Four Phases Of Forensic Comparison

In order to draw meaningful conclusions from a comparison between physical items of evidence, a forensic scientist generally must engage in a process with four phases:

**Phase 1:** Sample evaluations and analyses;

**Phase 2:** Comparison of samples (“matching” or “grouping”);

**Phase 3:** Assessing the *likelihood* (relative frequency) of the matching features in the relevant population from which the evidence originated;

**Phase 4:** Conclusions about whether the samples have (or are likely to have) a common source.

Just as a weak link can compromise the strength of a chain, weakness at any phase of this process can undermine the reliability of the expert’s ultimate conclusions.

In Phase 1 (evaluation and analy-

sis), the analyst assesses the physical, chemical, and/or mechanical characteristics of the samples. For many types of evidence such as latent prints, bitemarks, and toolmarks this assessment involves a simple visual examination, sometimes aided by a magnifying glass or microscope. For other types of evidence, such as biological and chemical samples, the assessment may begin with a visual examination and then proceed to the use of sophisticated analytic instruments to generate more detailed information about the composition or other characteristics of the samples. Forensic scientists use a variety of

---

“It is not sufficient simply to have a second analyst ‘confirm’ the conclusions of the first analyst if neither analyst is blind to the expected result. Both analysts are likely to suffer from the same tendency to confirm what is expected.”

---

impressive instruments, including nuclear reactors, genetic analyzers, and complex radiation sensing and electronic counting equipment. It is a mistake, however, to assume that use of a sophisticated instrument will necessarily lead to valid conclusions. The error in this thinking is nicely illustrated by CBLA evidence, where FBI analysts used magnificent and enormously expensive instruments such as a nuclear reactor and later a technique known as inductively coupled plasma atomic- (or optical-) emission spectrometers, to generate conclusions that were meaningless.

In Phase 2 (comparison), the analyst compares the analytical results for “questioned” and “known” evidentiary samples with each other, and sometimes with known reference samples. When the analyst is trying to link a crime scene sample to a suspect, the analyst typically compares the analytical results for the crime scene sample to those from the suspect. For example, a latent print is compared to the suspect’s inked “ten print” card or the DNA profile of a

bloodstain is compared to the DNA profile of a blood sample from the suspect. When the analyst is trying to identify the nature of the sample — whether white powder seized from a suspect is cocaine, for example, or whether particles from the suspect’s hand are gunshot residue (GSR) — the analyst will typically compare the analytical results for the evidentiary sample to the results for known standards representing the substance in question. In either case, the analyst is trying to determine whether the evidentiary sample can be distinguished from the “known” reference sample or standard. If the analytical results are sufficiently different, then the analyst declares the items “analytically distinguishable” (this conclusion is sometimes characterized as a “non-match” or “exclusion”). On the other hand, if the analyst cannot rule out the possibility of a common source, the analyst will deem the results to be “analytically indistinguishable.” Forensic scientists sometimes use the term “match” or “inclusion” to describe this conclusion. When dealing with more than a few samples, they may divide them into groups such that items within each group are deemed “analytically indistinguishable” from each other.

One problem analysts face when deciding whether items are “analytically distinguishable” is that the history of each item may unknowingly affect the analytical results. Aggressively cleaning a gun may alter the markings that the gun makes on bullets, for example, such that the striation pattern on an evidentiary bullet may differ somewhat from the pattern on a reference bullet fired from the same gun. Exposing a biological sample to the environment may degrade it in ways that cause its DNA profiles to differ from the profile of a reference sample from the same person. Thus, analysts must make allowance for some differences in analytical results in order to avoid mistakenly “excluding” samples or worse, rendering a “false positive” conclusion based on marks imparted under different conditions or from different sources.

When using analytical instruments to make comparisons, analysts also face the related problem of measurement error. Analytical testing is always subject to a degree of imprecision or measurement error. If you test the same sample twice, you won’t get exactly the same reading — there is a margin of error that must also be taken into account when deciding whether two unknown samples are “analytically indistinguishable.”

In order to avoid “false negatives” —

that is, instances in which samples with a common source are mistakenly called a “non-match” or “exclusion,” analysts must make allowance for differences in analytical results produced by the different history of the samples and by measurement error. If they make too many allowances, however, they greatly increase the chances of a “false positive” — i.e., “matching” or “grouping” samples or marks from different sources. Ideally, the analyst will set the threshold or standard for declaring a “match” (hereafter “match criterion”) in a manner that minimizes the likelihood of false negatives and false positives. Whether the match criterion employed in a particular case was appropriate is always an issue that warrants close scrutiny. The use of sophisticated instruments does not guarantee that analysts will use sophisticated or reasonable match criteria, regardless of the precision of the instrumentation. Nor is it always the case that the match criterion is part of an established *and written* protocol. Even when there is a written protocol that describes the criterion used to claim a match, it is not always based on reliable scientific studies and may, as was the case for CBLA, be completely arbitrary and subjective. Whether the match criteria that a laboratory is applying makes sense or not in light of available research is sometimes a difficult question that can only be evaluated by a statistician or a scientist with overlapping expertise in statistics.

For some types of forensic analyses, such as handwriting, bitemark and toolmark examination, the judgment that items “match” or are “analytically indistinguishable” appears to rest entirely on the analyst’s subjective judgment.<sup>6</sup> The only “standards” these fields claim to apply are, on close examination, merely vague statements amenable to almost any interpretation that the analyst chooses to give them. To make matters worse, analysts often fail to record, photographically or otherwise, what features or characteristics led them to conclude that two samples “match” or do not “match.” Hence, the match determination is difficult or impossible to assess and rests entirely on the *ipse dixit* of the expert.

In Phase 3 (assessing *likelihood* or relative frequency of the matching characteristics), the forensic analyst either explicitly or implicitly assesses the likelihood of the match. In other words, the analyst assesses the chance of a coincidental match in the relevant population of similar items. Suppose, for example, that two bloodstains are analyzed and

each is found to contain genetic factor X. In order to assess the value of this match, the analyst must obviously consider the probability of finding factor X in both samples if the samples came from different people. This probability will depend on the rarity of factor X. If factor X is found in one person in a billion the “match” is far more significant than if factor X is found in one person in 10. And, of course, this match means nothing if everyone has factor X. So without knowing the rarity of factor X, it is impossible to draw any meaningful conclusions about the likelihood the samples had a common source from evidence about the “match” based on factor X. In the case of a bullet composition match, the match obviously means nothing if everyone who owns bullets in the locality of the crime owns bullets of similar composition.

When assessing the overall value of a “match” implying common source, the analyst should consider two factors: first, the probability of obtaining the observed results if the samples have the same source; and second, the probability of obtaining the observed results if the samples have different sources. The probative value of the forensic evidence (for proving the samples have the same source) depends on the extent to which the first probability exceeds the second.<sup>7</sup> Consequently, accurate assessment of these probabilities is a crucial foundational element for *any* statement that *any* forensic scientists might make about the meaning of *any* forensic “match.”

In Phase 4, which we call the inference/conclusion phase, the analyst draws explicit conclusions about the meaning of the claimed “match” — such as whether it means that the items had (or probably had) a common source. With CBLA evidence, FBI analysts often testified that “analytically indistinguishable” bullets came from (or likely came from) the same box of bullets or from boxes manufactured in the same facility on the same day. With fingerprint and toolmark evidence, analysts often testify that “analytically indistinguishable” marks are certain to have been made by the same finger or same tool to the exclusion of all other fingers (or tools) in the world. With DNA evidence, analysts sometimes testify “to a scientific certainty” that samples with matching profiles came from the same person. The logic underlying such inferences is often questionable and demands close examination. All too often forensic analysts leap from the observation of “same characteristics” to the conclusion of “same source” without

any convincing evidence of the connection. The “same composition = same source” fallacy passed muster in the courtroom for decades in the case of CBLA evidence before it was recognized as an absurdity. In other areas of forensic identification science, similar forms of fallacious logic have yet to be challenged.

## Evaluating A Case

The most important step in preparation for facing the forensic expert is to *request a curriculum vitae and all bench notes, underlying data, and demonstrative exhibits* of the prospective expert as early as possible. Some of the data resulting from the sophisticated instrumental analyses used today can require days to review, graph, and evaluate. In many jurisdictions defense counsel are routinely provided only one- or two-line reports on forensic examinations and conclusions; nothing more. It may have been standard practice in the past to proceed to trial based on such reports without examining the underlying laboratory notes and data. It may now constitute ineffective assistance of counsel, or even malpractice. To effectively represent a client who is incriminated by forensic identification evidence, you must get the bench notes, study them, and understand exactly what was done. Then you must evaluate each of the four phases of the forensic comparison process and identify any scientific weaknesses or flaws.

### *Evaluation of Phase One (Analysis)*

When evaluating Phase One (sample evaluation and analysis), it is crucial to know exactly which features of the samples were examined or considered in the analysis (and which features were not considered). If the analyst looked at a latent print, which features of the print were identified as significant or probative and which features were disregarded? If the analyst looked at striations on a bullet, what was the exact pattern that was seen and where were the striae observed? If an instrument was used to analyze the samples, what exactly did it measure (did it capture a significantly relevant characteristic?), how much of the sample was characterized (was it “representative” of the bulk?), where on the sample was it measured, and what were the specific findings? Knowing the specific features that were considered in the analysis is a necessary first step toward assessing the match criteria that were applied (Phase 2), the likelihood of a coincidental match (Phase 3), and the ultimate conclusions drawn from the analysis (Phase 4).

It is important to consider whether

any instrument being used in the analysis is appropriate for that specific purpose. Do not make the mistake of assuming that an instrument that is useful and reliable for another purpose is necessarily appropriate for the specific forensic analysis you are evaluating. While it may seem absurd to think that defense lawyers should be responsible for evaluating the appropriateness of the instrumentation employed in a forensic lab, history has shown the necessity of asking this basic question. The technique of choice for comparative bullet lead analysis for over 25 years (from the early 1960s until about 1995) was neutron activation analysis (NAA), which required use of a nuclear reactor. Although this was sophisticated high-tech analytical instrumentation at its finest, it was *inappropriate* for the conclusions rendered in bullet lead cases because it could only effectively analyze for three elements (analytes): antimony, copper and arsenic, and could not sufficiently discriminate compositions within the most common range for bullets.<sup>8</sup>

The use of NAA for analysis of bullet lead illustrates two key points. First, it shows the longstanding tendency of forensic scientists to present misleading testimony based on inadequately validated methods. Second, it shows the importance of knowing exactly what is being tested in a forensic assay. For years, FBI examiners got away with asserting that their test produced a chemical profile of bullets that could identify with great accuracy bullets that originated in the same molten source of lead, same manufacturer, or even the same box. As it turned out, they were wrong. Yet it is only by considering exactly what the assay was testing (levels of antimony, copper and arsenic only) that one can begin to ask sophisticated questions that can identify flaws in the evidence — like how the FBI analysts decided that two samples were similar enough in the levels of these elements to be deemed a “match” (Phase 2), how likely such a “match” might be if the bullet came from different molten sources, different manufacturers, or different boxes (Phase 3), and whether a “match” really means that the bullets in question are likely to be from the same box (Phase 4).

After verifying that the instrument used was appropriate, it is also important to check that the instrument was properly calibrated, that proper control samples were run, and that the instrument produced the expected results for all control samples. Typically laboratories will run “positive controls” (samples of known

composition) to demonstrate that the instrument is capable of producing accurate results (rather than “false negatives”). Labs also run “negative controls” (sometimes called “blanks”) that are known not to contain the element or characteristic the instrument is designed to detect in order to demonstrate that the instrument is not giving “false positive” readings. If the controls produce unexpected results, it indicates something is wrong with the procedure and therefore invalidates any conclusions drawn from it. Hence, any hint of problems on control samples warrants close attention. If the laboratory fails to provide complete information about controls, the reported results should be viewed with suspicion.

It is also important to identify any instances in which the analyst relied on subjective judgment to modify or override instrumental results. In DNA analysis, for example, analysts can use software to modify the results produced by genetic analyzers. If the analyst thinks a particular “allele” (genetic marker) detected by the instrument is spurious, the analyst can simply delete it. In bullet lead comparisons, analysts frequently ignored one of the critical analytes (copper) if copper didn’t match, asserting probable jacket contamination. Whether such deletions are appropriate or not is, obviously, impossible to assess if you do not know that it occurred. Although a competent, conscientious analyst will document any “overrides” in laboratory notes, many analysts are not so careful. In some instances operator interventions to “modify” test results become apparent only when an independent expert is allowed to reanalyze the original instrumental data. Fortunately, many modern analytical instruments store a copy of the original instrumental output (often called “raw data”) as a computer file. Obtaining copies of these files and having them reanalyzed can sometimes be very revealing.<sup>9</sup> Labs that refuse to disclose such files (or claim that they discard such files routinely) should be viewed with suspicion. There is no good reason for destroying the scientific data underlying a forensic analysis or for refusing to disclose it to defense counsel.

*Evaluation of Phase Two (Comparison)*

When assessing the comparison of samples, it is important first to satisfy yourself that appropriate samples are being compared. If a forensic expert is comparing soil found on a suspect’s boot to soil from a crime scene, was the comparison soil collected from the right



**FEDERAL CIRCUIT ORAL ARGUMENT PANEL**

NACDL members with an oral argument scheduled in a federal court of appeal can hone their skills before a panel of mock judges consisting of attorneys, professionals, and law students. At least six weeks prior to oral argument, members wishing to use this service should submit a one-page “bench memo” to the coordinator(s) listed below in the circuit where the case will be heard. Coordinators will screen cases and arrange for a panel of mock judges. The panel will attempt to accommodate every request, but given the volume of federal appeals, the coordinators will have the discretion to prioritize cases, favoring those that raise significant appellate issues.

**CIRCUIT COORDINATORS**

Please contact the coordinator in the circuit where the case will be heard.

**PANEL CO-CHAIRS**

<b>Howard M. Srebnick</b> Black, Srebnick & Komsipan, PA 201 S. Biscayne Blvd., Ste. 1300 Miami, FL 33131 305-371-6421 Fax 305-358-2006 hsrebnick@royblack.com	<b>Scott A. Srebnick</b> 2400 S. Dixie Hwy. Ste. 200 Miami, FL 33133 305-285-9019 Fax 305-854-8782 srebnick@aol.com
--	---

<b>1st Circuit</b> <b>Kim Homan</b> Boston, MA 617-227-8616 homanlaw@aol.com	<b>7th Circuit</b> <b>Allan Ackerman</b> Chicago, IL 312-332-2891 doctrrips@aol.com
<b>Rachel Brill</b> Hato Rey, PR 787-753-6131 rabrill@attglobal.net	<b>8th Circuit</b> <b>Grant Shostak</b> St. Louis, MO 314-725-3200 gshostak @msmattorneys.com
<b>2nd Circuit</b> <b>Susan Necheles</b> New York, NY 212-997-7595 srn@hafetzlaw.com	<b>9th Circuit</b> <b>Shereen Charlick</b> Miami, FL 305-530-7000 shereen_charlick@fd.org
<b>3rd Circuit</b> <b>Ted Simon</b> Philadelphia, PA 215-563-5550 tsimonesq1@aol.com	<b>Martin Sabelli</b> San Francisco, CA 415-575-8824 piropo@msn.com
<b>4th Circuit</b> <b>Terry Reed</b> Alexandria, VA 703-299-8734 tgreed@lcrfirm.com	<b>10th Circuit</b> <b>Norman Mueller</b> Denver, CO 303-831-7364 nrmueller@hmflaw.com
<b>5th Circuit</b> <b>Jancy Hoeffel</b> New Orleans, LA 504-865-5895 jhoeffel@law.tulane.edu	<b>11th Circuit</b> <b>Ellen Podgor</b> Gulfport, FL 727-562-7348 epodgor@stetson.edu
<b>William Kim Wade</b> Dallas, TX 972-233-1100 kwade@wadelaw.com	<b>Scott A. Srebnick</b> Miami, FL 305-285-9019 srebnick@aol.com
<b>6th Circuit</b> <b>John Feldmeier</b> Cincinnati, OH 513-721-4876 feldmeierj@aol.com	<b>DC Circuit</b> <b>Neil Jaffe</b> Washington, DC 202-208-7500

places? In one such comparison known to the first author, the reference soil samples were obtained from six inches below the surface of the ground (supposedly to preclude surface contaminants). How many suspects flee a crime by stepping on soil six inches below the surface of the ground? The effort to obtain an uncontaminated reference sample produced an unrepresentative (and possibly meaningless) reference sample.

In cases where an evidentiary sample was exposed to conditions that may have changed its properties, such as a fire, explosion or exposure to a harsh environment, forensic analysts sometimes find it helpful to expose reference samples to similar conditions before making the comparison. However, exact duplication of the relevant conditions is often impossible (even if the conditions are known) and seemingly insignificant changes in the simulated conditions often have dramatic effects on the subsequent comparisons. Consequently, any forensic analysis that involves “simulation” of “relevant conditions” warrants close scrutiny.

Second, consider whether the analyst used an appropriate criterion (match standard) to reach conclusions about “matches,” “positive associations,” or “inclusions.” If the laboratory is using a quantitative standard for declaring a “match,” are there studies to support that particular threshold or cut-off used? Were those studies conducted under conditions similar to those that apply in the case at hand? And what do those studies say about the probability of a false inclusion or a false exclusion under the relevant conditions? If you cannot answer these questions, you need to dig deeper into the underlying scientific validation for the match standard.

If the forensic analyst declared a “match” based on subjective criteria (as typically happens in latent print and toolmark examinations) what evidence is there that such subjective judgments can be made reliably — that is, that different experts would agree regarding what constitutes a “match” and what does not? And more importantly, what evidence is there that the expert’s subjective determinations are accurate? Is there any research on this expert’s (and any other experts’) false positive and false negative rate when evaluating similar samples? If the standard for declaring a match is simply the say-so of the expert, is there any evidence to support the claim that this expert can draw such conclusions accurately? Without very good and convincing answers to these questions, the

subjective judgment of the analyst is impossible to evaluate.

A key issue is whether a sufficient number of characteristics have been examined in the effort to distinguish the samples. Items that appear to be identical in all physical and even chemical aspects can be quite different as to mechanical properties. In such a case, if it walks like a duck, quacks like a duck, and looks like a duck, it may still not be a duck. Opinions about “matches” or positive associations can be quite vulnerable to error if too few, or non-meaningful, properties or characteristics are compared. The FBI’s practice of examining only three analytes when using NAA to compare bullets is an excellent example of this problem.

---

**“History has shown that assertions of expertise, unsupported by scientific research, are often wrong.”**

---

Third, consider whether the analyst took adequate measures to control for “observer effects” or “confirmation bias” — that is, for the human tendency to see what one expects and/or desires to see.<sup>10</sup> When an analyst analyzes and evaluates an evidence sample, it is all too easy to be influenced by knowledge of the reference sample, particularly if the analyst is aware of other evidence regarding whether the samples should “match.” When FBI fingerprint examiners evaluated the notorious fingerprint from the Spanish train bombing case, for example, they were apparently influenced by their knowledge of the details of the reference print from the suspect — Oregon attorney Brandon Mayfield.<sup>11</sup> Sections of the crime scene print that were similar to Mayfield’s were credited as reliable data. Sections of the evidence print that differed from Mayfield’s were dismissed as the product of distortion in the underlying surface or an overlay from another print. In this manner, the FBI examiners managed to incorrectly “match” Mayfield to a print that was not his. Spanish authorities determined that the print was actually that of an Algerian suspected terrorist named Ouhmane Daoud.

Observer effects have been the source of a great deal of scientific error and self-deception. In his iconic book *Galileo’s Revenge: Junk Science in the Courtroom* (1991), Peter Huber traced several false scientific theories to misin-

terpretation of data arising from uncontrolled observer effects. A scientist who is committed to a pet theory inevitably (and unconsciously) interprets data in a manner consistent with that theory. By Huber’s account, uncontrolled observer effects are one of the hallmarks of junk science. Unfortunately, this hallmark of junk science is “a rampant and uncontrolled part of normal practice” in the field of forensic science.<sup>12</sup> While academic scientists generally take careful steps to control for observer effects, such as conducting studies in a “blind” or even “double-blind” manner, forensic scientists almost always fail to do so.

Observer effects are most likely to introduce bias and distortion when analysts are relying on subjective judgment to decide which features of the data are important and whether the samples are sufficiently similar to constitute a “match.” Hence, observer effects are likely to be particularly influential for fingerprint and toolmark evaluations, which rely almost entirely on subjective determinations (and often are inadequately documented as well). However, subjective judgment can play an important role even in the interpretation of seemingly more objective instrument-based tests like DNA and CBLA analysis, so the potential for observer effects is important to consider there as well.<sup>13</sup>

Forensic scientists can generally control for observer effects if they care to do so. Evidentiary samples should be evaluated and analyzed while the analyst is “blind” to the features of reference samples. For example, the FBI fingerprint examiners should have decided which features of the Spanish train print constituted reliable data, and which were distorted or overlaid, before knowing whether the features were consistent or inconsistent with Mayfield’s fingerprint. Similarly, DNA analysts should decide which “alleles” in an evidentiary sample profile are real and which are spurious before knowing whether those alleles match up with a suspect’s. Analysts who are “blind” to the consequences of the “match” determination should generally be the ones who make the comparison.

One approach to blind interpretation is to separate the various parts of the analysis and assign them to different people. The evidentiary and reference samples could be analyzed and evaluated by different analysts. A third person, who knows nothing about the case, could decide whether the samples “match.” Another approach is simply to perform the analysis and evaluation of various samples in a sequence so that analysis

and interpretation of early samples cannot be influenced by knowledge of the later samples. Evidentiary samples, which are generally more difficult to evaluate, can be analyzed first, before the analyst knows the features of the reference samples. If the analysis of the evidentiary sample is committed to writing before the analyst knows the features of the reference sample, then observer effects cannot influence the analysis of the evidentiary samples.

When evaluating forensic evidence, it is important to know whether simple, obvious steps to reduce observer effects, such as these, have been taken. Failure to control for observer effects may affect the reliability of the findings in ways that arguably go to their admissibility under *Daubert* rather than always going to their weight.

It is not sufficient simply to have a second analyst “confirm” the conclusions of the first analyst if neither analyst is blind to the expected result. Both analysts are likely to suffer from the same tendency to confirm what is expected. If the second analyst knows about the conclusions of the first analyst, the situation is even less satisfactory. The FBI’s error in linking Brandon Mayfield to the Spanish train bombing again illustrates the problem. After a prominent FBI fingerprint examiner made the initial incorrect match, two additional FBI experts examined the prints and “confirmed” the false match. A fourth examiner, hired by the defendant, also reviewed and “confirmed” the false match. Tellingly, a report issued by the FBI after the error came to light attributes the incorrect confirmatory opinions to “confirmation bias.” “[B]ecause the initial examiner was a highly respected supervisor with many years of experience,” the report states, “it was concluded that subsequent examinations were incomplete and inaccurate. To disagree was not an expected response.”<sup>14</sup> The FBI report endorses blind verification, which would be a significant improvement over current practices. Needless to say, if blind verification is necessary to produce accurate fingerprint identifications, it is likely also to be necessary to assure the accuracy of other types of forensic identifications.

#### *Evaluation of Phase Three (Frequency of Features)*

Assuming you are satisfied that the incriminating specimens were properly characterized as indistinguishable (therefore a “match” or “inclusion”), the next step for assessment of forensic significance involves estimation of probabilities

for determination of probative value. As noted earlier, there are two crucial questions: (1) how likely are the observed results if the samples had a common source; and (2) how likely are the observed results if the samples did *not* have a common source? Without valid answers to both questions, there is no way to assess the probative value of the forensic evidence for proving that the “matching” items had a common source. Yet there is tremendous variation in how well forensic disciplines have validated this phase of the process.

For example, there is substantial scientific literature that documents the frequency of the genetic markers used for DNA matching in a variety of human populations. So when a DNA analyst says STR profiles are unlikely to match if the samples are from different people, there is (usually) a solid scientific foundation for that conclusion.<sup>15</sup> For most other types of forensic identification evidence, however, the scientific literature is minimal or non-existent on the frequency of the characteristics that are examined when “matching” samples. Analysts who testify about toolmarks, bitemarks, handwriting characteristics, latent prints, bullet lead compositions and the like are usually relying solely on their own implicit assessment of the rarity of “matching” features.

When pressed to explain how they know that the samples have a common source or unlikely that the samples have a different source, analysts frequently resort to claims of support from “my lengthy experience” or “my knowledge, training and expertise.” In the absence of published studies that address the relevant questions, any claims about the likelihood of these events should be viewed with skepticism. History has shown that assertions of expertise, unsupported by scientific research, are often wrong. For years FBI analysts claimed to “know” that a bullet lead match was highly unlikely if the bullets came from a different box or molten source. In fact, as the FBI Laboratory Director has now acknowledged, their claims to such knowledge were groundless. Until tested and confirmed by meaningful and comprehensive scientific studies, any claims that a forensic scientist makes about the rarity of a particular mark or characteristic in a particular population should be viewed as speculation rather than science.

Although this point may seem obvious, it has apparently escaped the notice of lawyers and judges in many cases. In 1997, a judge in a South Carolina murder case questioned how meaningful a bullet lead “match” might be — wondering whether “everybody in town” had

## NACDL Member Lapel Pin

Only \$20.00



Order on-line at  
<http://www.nacdl.org/onlineshop>

**Now Affordably Yours!**

bullets of the same composition.<sup>16</sup> Although the question is obvious, it was rarely asked in other cases. The most surprising aspect of the incident is that the FBI did not and could not answer the question — no one knows how common “matching” bullets might be in a particular town. While the FBI was willing to assume that a compositional match is a rare event, the frequency of matching bullets in any particular geographical locality was never actually studied until recently. Data now emerging confirm intuitive expectations and the South Carolina judge’s concern that bullets may be distributed in a manner that creates regional concentrations of bullets with “matching” compositions — a possibility the FBI never bothered to check and that defense lawyers raised in only a few isolated cases.<sup>17</sup>

When forensic scientists cite studies to back up their claims about the likelihood of the results, you should scrutinize the studies carefully and ask hard questions. What is the source of the samples in the study? What process was used to select samples for inclusion in the study? Are the samples representative of a relevant population (for example, how relevant are bullet compositions from Los Angeles in a Detroit murder trial)? Do the statistical estimates derived from the study depend on any assumptions, such as the assumption that different characteristics are independent? If so, is there any evidence that those assumptions are valid?

In forensic science there is a long history of questionable and misleading empirical research on statistical issues. These studies often emerge just as the claims of forensic scientists are being challenged. In response to challenges to the reliability of fingerprint identification, for example, FBI employees Stephen Meager, Bruce Budowle, and David Ziesig in 1999 produced their notorious, unpublished “50K” study in which 50,000 fingerprint images were each compared to all the others. Although this study was cited repeatedly in courtroom testimony as evidence that the probability of a false fingerprint match is infinitesimally small, academic critics have excoriated the study as a piece of scientific trash.<sup>18</sup> One prominent expert called it “extraordinarily flawed and highly misleading.”<sup>19</sup> Two prominent European experts said the study “so transcends reality that we are amazed that it was admitted into evidence. It is entirely unsupportable.”<sup>20</sup> The basic problem was that the study failed to take into account any variation

in prints made by the same finger, so that its statistical conclusions represented an idealized and unrealistic situation with no connection to reality.

Similarly, when challenges were raised to CBLA evidence, the FBI came up with another misleading study that purported to prove that the probability of a CBLA match between *unrelated* bullets is extremely low.<sup>21</sup> This study illustrates a number of common problems with forensic “databases.” The samples in the FBI’s “database” of bullets were not selected randomly from the population of bullets available commercially at a given point in time in any given locale. Instead the researchers selected bullets from casework that had been performed in the FBI lab over a period of many years using several criteria apparently designed to assure the greatest possible diversity in the samples (and thereby reduce the likelihood of coincidental matches among the database samples).

For example, the researchers attempted to assure that each bullet in the “database” came from a distinct molten source of lead. Yet manufacturers make thousands or millions of bullets from the same molten source and those bullets end up in different boxes of ammunition that are frequently distributed in the same geographic area.

To make matters worse, the researchers threw together bullets of many different calibers and types. This further increased the diversity of the database (and reduced the likelihood of coincidental matches) because there are different metallurgical considerations underlying manufacture of the different types of bullets. A bullet from a .22 caliber handgun, for example, is unlikely to match shot lead and less likely to match a bullet from a .50 caliber military weapon. However, because casework comparisons typically involve bullets of the same caliber and type, the relevant issue at trial is the probability of a random match *within* a particular caliber and type of bullet (e.g., the probability of a match between two .22 caliber handgun bullets) rather than the probability of a match in the database as a whole comprising bullets from distant geographical regions.

Thus, by constructing a “database” of unrepresentative samples and using it to address irrelevant and inappropriate questions, the FBI researchers created the false impression that coincidental matches in casework are rare. Fortunately, the problems in this FBI study have now been widely recognized,<sup>22</sup> but it is unlikely to be the last

misleading study that makes its way into the criminal courts.

#### *Evaluation of Phase 4: Inferences and Conclusions*

After completing the first three phases of the forensic process, the examiner reaches the inference/conclusion phase (Phase 4). At this stage the examiner prepares an opinion (inference or conclusion) for the contributing agency and possibly court. The work at this stage is the tip of the forensic iceberg — the portion most readily seen — but it depends on the work done in the previous stages. The conclusion is the culmination of a chain of inferences and is only as strong as the weakest link in that chain.

When evaluating forensic evidence, it is often helpful to create a list of the inferences that the analyst made to reach the ultimate conclusion. The inferences can usually be characterized as a series of “if-then” statements. Laying out the inferences in detail makes it easier to see weaknesses in the final conclusions. For example, the chain of inferences underlying CBLA might look like the following:

(1) IF [the instrument is properly calibrated; the controls performed as expected; the results were recorded properly, etc.] THEN [the bullet from the crime scene has the chemical profile I have specified for it; the reference bullets from the defendant’s box have the profiles I have specified for them].

(2) IF [the bullet from the crime scene has the chemical profile I have specified for it; the reference bullets from the defendant’s box have the profiles I have specified for them] THEN [the crime scene bullet is “analytically indistinguishable” from the bullets in defendant’s box].

(3) IF [the crime scene bullet is “analytically indistinguishable” from the bullets in defendant’s box] THEN [the bullets are very likely to have originated in the same molten source of lead at a manufacturer].

(4) IF [the bullets all originated from the same molten source of lead at a manufacturer] THEN [the bullets are all likely to have come from the same box].

As this example shows, the ultimate

conclusion depends on a chain of “if-then” inferences. When evaluating a forensic opinion, you should be sure that each “if” step used as a foundational stepping stone to the final inference is valid, then verify that every progression *between* each “if” and “then” in the logic process is valid and supportable. In the case of CBLA, there were weaknesses at several stages. Although the instrumentation the FBI used was appropriate (once the FBI made the shift from NAA to more modern inductively coupled plasma emission spectrometers), the National Research Counsel strongly criticized the “match criterion” the FBI used for declaring samples to be “analytically indistinguishable,” saying it was too lenient and created too great a danger of false positives. Moreover, even if bullets are “analytically indistinguishable” there are no comprehensive or meaningful foundational studies (“no body of data” as observed by the *Mikos* court<sup>23</sup>) to support the next inference — that “indistinguishable” bullets likely originated in the same molten source of lead. Finally, and perhaps most importantly, even if bullets originated in the same molten source of lead, there is no way of knowing how probative that fact is for showing they came from the same box, given the potential for large numbers of bullets from the same molten source to appear in different boxes distributed in the same geographic area.

In Phase 4, forensic analysts often make the leap from thinking about the probability of the observed results under various hypotheses (which occurs in Phase 3) to drawing conclusions about the probability that a particular hypothesis is true. For example, a fingerprint (or toolmark) examiner will typically move from a probabilistic assessment — *e.g.*, “it is very unlikely that I would see so many similar details if these two prints (marks) were made by different fingers (tools)” — to a conclusion, *e.g.*, “these prints (marks) were made by the very same finger (tool) to the exclusion of all other fingers (tools) in the world.”

You should always be skeptical when a forensic scientist claims to have identified an item uniquely as originating from one and only one possible source in all the world. These claims rest implicitly on the assumption that the “matching” characteristics are so rare as to be unique — an assumption that is impossible to test or verify. Even when dealing with DNA evidence, where statistical estimates of rarity are well supported, it is problematic for an analyst to claim to have identified the source of a sample

uniquely as coming from one and only one person in the universe. Such claims rest not on science but on an arbitrary conclusion that the “matching” genetic characteristics are *sufficiently* rare to rule out the chance that anyone else could have a duplicate DNA profile. When dealing with fingerprints, bitemarks, bullet lead compositions or toolmarks, where there is little or no scientific basis for any statistical estimates of rarity, it is even more problematic for an analyst to claim to have identified the source of a mark or bullet uniquely (although analysts in these areas have persisted in doing so).

You should also be skeptical, and should raise objections, when a forensic scientist expresses an opinion about the probability or likelihood of any ultimate conclusion. For example, while it is generally permissible for a DNA expert to testify about the probability that an evidentiary sample would happen to match the defendant *if* it actually came from a random person other than the defendant, it is improper for the DNA expert to testify directly about the probability that the evidentiary sample *came from the defendant*. (If you do not understand the distinction just drawn, please read the previous sentences again, carefully). The difference between the permissible and impermissible statement is the nature of the foundation. There often is a scientific foundation (population studies) for estimating the likelihood of finding a particular DNA profile in a person randomly drawn from the population. However, there is not and can never be a *scientific* foundation for estimating the probability an evidentiary sample came from the defendant because that probability depends on a variety of factors (*e.g.*, the defendant’s alibi, other evidence in the case) that have nothing to do with scientific evidence. The forensic expert is in no position to evaluate the defendant’s alibi and other evidence, and generally has no business basing a conclusion on such factors. So, as a matter of basic evidentiary principle, such conclusions are generally improper. You should object to such conclusions on grounds that they lack proper foundation, that they invade the province of the jury, and that they go beyond the expert’s competence.<sup>24</sup>

**Can Forensic Testing Be Validated By Proficiency Testing?**

Often the only research that forensic analysts can muster to support their claims of accuracy in Phase One and Phase Two is “proficiency testing.”<sup>25</sup> Yet proficiency testing in forensic science is

**Membership Directory Changes for the 2007 Membership Handbook are due before September 1, 2006**

Please use this form to notify NACDL of any changes to your name, address, area code, phone number, fax number, or e-mail address. Be sure to include your new phone and fax numbers when you submit an address change.

**OLD Address:** PLEASE PRINT CLEARLY

Name: \_\_\_\_\_  
 Address: \_\_\_\_\_  
 City: \_\_\_\_\_  
 State: \_\_\_\_\_ Zip: \_\_\_\_\_  
 Country: \_\_\_\_\_  
 Phone: \_\_\_\_\_  
 Fax: \_\_\_\_\_  
 E-mail: \_\_\_\_\_

**NEW Address:**

Name: \_\_\_\_\_  
 Address: \_\_\_\_\_  
 City: \_\_\_\_\_  
 State: \_\_\_\_\_ Zip: \_\_\_\_\_  
 Country: \_\_\_\_\_  
 Phone: \_\_\_\_\_  
 Fax: \_\_\_\_\_  
 E-mail: \_\_\_\_\_

Please complete form and fax to 202-872-8690 or E-mail Mary Ann Robertson at mar@nacdl.org or call 202-872-8600 x222

frequently worthless as a true indicator of examiner proficiency. It is often designed as window dressing to create apparent support for the laboratory's claim of competence without taking the risk of seriously testing that claim. When evaluating proficiency testing, it is important to consider several factors.

Consider first the basic structure of the test. The first author has observed proficiency testing (not necessarily at the author's former law enforcement laboratory) where the supervisor repeatedly (over a period of years) selected the samples to be tested from a collection of only 4 or 5 samples! In that situation, analysts became so intimately familiar with the "proficiency samples" that they didn't even need to run a complete analysis to identify them. The second author has reviewed DNA proficiency tests in which the "correct" answer in every instance was that the samples "matched." Consider how likely an analyst would be to make a false match in that test. Consider further what it means for the analyst to announce, after taking a series of such tests, that he has never, ever had a false match.

Another consideration is the nature of the samples — are they "real world" specimens that are representative of the samples actually tested in your case, or are they from Fantasyland? In one instance known to the first author, a lab tried to validate its procedures for identifying components of an exploded bomb by doing proficiency tests on similar items obtained from retail shelves that had not been in an explosion, even though one of the elemental constituents that comprised the bomb components was particularly vulnerable to volatilization at the elevated temperatures existing during an explosion.

A key question is: what does the proficiency test test? Many proficiency tests focus on just one or two phases of the four-phase process of forensic inference outlined in this article. Asking a bullet lead analyst to measure the level of various elements in samples of lead alloy provided by NIST (National Institute of Standards and Technology) may well be a good test of the analyst's proficiency at Phase One of the process, but it does not validate CBLA as a whole. "Proficiency tests" that focus on uncontroversial phases of a forensic test, such as the first phase in CBLA analysis, while ignoring other more controversial phases, are not particularly helpful and often seem designed to provide meaningless assurance about a test rather than seriously examine its accuracy.

Another issue is whether the proficiency test has been conducted and reported fairly. Although chicanery in proficiency testing is hard to prove, there are strong incentives for "wink and a nod" testing. There is serious potential for embarrassment and loss of credibility, for the analyst and for the lab, if an error is reported. This is an area in which the strong norms of camaraderie and solidarity that exist in the law enforcement community may work against objective scientific endeavor.

Another way to virtually eliminate the possibility of proficiency failures is to devise tests that examiners can pass without much deliberation or effort. When Allan Bayle, former Scotland Yard fingerprint examiner, evaluated the FBI's internal proficiency tests of its fingerprint examiners, he concluded they were too easy to be meaningful. Bayle testified, "[I]f I gave my experts these tests, they'd fall about laughing," and characterized the FBI proficiency tests as "a joke."<sup>26</sup> Unfortunately, it is usually difficult to evaluate proficiency tests in any objective way due to agency secrecy and limitations on discovery.

## Conclusion

The ancient Chinese had a famous curse: "may you live in interesting times." These are indeed interesting times for the forensic sciences. The field has been glorified in the media and for much of its history has received a free pass into the courtroom. But the times they are a-changing. In the academic community there is growing awareness that many areas of forensic science are problematic — undervalued and oversold. And that spells big trouble for forensic science. We are entering a weeding out period in which weaker techniques like CBLA will either improve or be abandoned, and exaggerated claims will either be moderated or excluded from testimony.

Lawyers who represent criminal defendants will play a crucial role in this process. In criminal cases, the junk science being presented in the courts is overwhelmingly prosecution junk science. Hence, the job of exposing it and challenging it, so it can be weeded out or improved, falls primarily to defense lawyers. The job is difficult but it can and must be done. We hope the framework provided in this article makes the task a bit easier.

## Notes

1. See e.g., Simon A. Cole, *More Than Zero: Accounting for Error in Latent Fingerprint Identification*, 95 J. CRIM. L. &

CRIMINOL. 985 (2005)(questioning validity and accuracy of fingerprint identification); Adina Schwartz, *A Systemic Challenge to the Reliability and Admissibility of Firearms and Toolmark Identification* 6 COLUMBIA SCI. & TECH.. L.REV., 1 (2005)(questioning validity of toolmark identification); Michele Nethercott & William C. Thompson, *Lessons from Baltimore's GSR Debate*. 29 THE CHAMPION 50 (2005)(exposing problems with gunshot residue evidence); Edward J. Imwinkelried & William A. Tobin, *Comparative bullet lead analysis (CBLA) evidence: Valid inference or ipse dixit?* 28 OKLA. CITY L. REV. 43 (2003)(critiquing bullet lead evidence); Michael J. Saks, *Banishing ipse dixit: The impact of Kumho Tire on forensic identification science*. 57 WASH. & LEE L. REV. 879 (2000); C.A. Stafford Smith & P. D. Goodman, *Forensic hair comparison analysis: Nineteenth century science or twentieth century snake oil?* 27 COLUMBIA HUM. RTS. L. REV. 227 (1996).

2. Michael J. Saks & Jonathan J. Koehler, *The coming paradigm shift in forensic identification science*. 309 SCIENCE 892 (2005).

3. See Imwinkelried & Tobin, *supra* note 1; NATIONAL RESEARCH COUNCIL, COMMITTEE ON SCIENTIFIC ASSESSMENT OF BULLET LEAD ELEMENTAL COMPOSITION COMPARISON, FORENSIC ANALYSIS: WEIGHING BULLET LEAD EVIDENCE 8 (2004)[hereinafter NRC REPORT]; William C. Thompson, *Analyzing the relevance and admissibility of bullet lead evidence: Did the NRC report miss the target?* 46 JURIMETRICS 65 (2005).

4. Charles Piller, *FBI Abandons Controversial Bullet-Matching Technique*, LOS ANGELES TIMES, Sept. 2, 2005, at A38.

5. Letter from Dwight Adams to Ralph Grunewald, Executive Director of the National Association of Criminal Defense Lawyers, Sept 1, 2005 (on file with authors).

6. Simon A. Cole, *Grandfathering Evidence: Fingerprint Admissibility Rulings from Jennings to Llera Plaza and Back Again*, 41 AM. CRIM. L. REV. 1189 (2004).

7. When evaluating forensic evidence, people sometimes focus solely on the probability of a random match—e.g., the probability of a DNA "match" if the samples came from different people. That can be a mistake because forensic analysts sometimes declare a "match" when they should not. For example, analysts sometimes declare a "match" between DNA profiles that do not completely correspond with one another (because analysts attribute the discrepancies to degradation or artifacts), see, William C. Thompson, et al., *Evaluating Forensic DNA Evidence: Essential Elements of a Competent Defense Review*. 27 THE CHAMPION 16 (2003). Because the value of the "match" for proving the samples have the same source is diminished to the

extent the "match" is itself an unlikely event if the samples have the same source, see Richard O. Lempert, *Modeling Relevance* 75 MICH.L.REV. 1021, 1025 (1977), failure to take this factor into account can lead to overvaluing forensic evidence.

8. In 1970, scientists from Gulf General Atomic reported to the then-Atomic Energy Commission (AEC) and the Law Enforcement Assistance Administration (LEAA) that the utility of using the nuclear technology for forensic purposes of comparing bullet compositions was severely limited for this reason. Courts, and presumably prosecutors, were realistically never apprised of those pioneering findings, yet examiners had testified for decades to claimed "matches" of bullets based on only two or three elements by NAA analysis.

9. For further discussion of the analysis of electronic data in DNA testing, see, Thompson, *et al.*, *supra* note 7; William C. Thompson, *Tarnish on the Gold Standard: Understanding Recent Problems in Forensic DNA Testing*, 30 THE CHAMPION 10 (2006).

10. See, Risinger, Saks, Thompson & Rosenthal, *The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion*, 90 CAL.L.REV 1 (2002)

11. See, W. Thompson & S. Cole, *Lessons from the Brandon Mayfield Case*, 29 THE CHAMPION 32 (2005)

12. Risinger, *et al.*, *supra* note 10, at 5.

13. See, W. Thompson *et al.*, *supra* note 7.

14. Robert B. Stacey, *A report on the erroneous fingerprint individualization in the Madrid Train Bombing Case*, 54 J.FORENSIC IDENTIFICATION 706 (2004).

15. Part of the reason that forensic experts have studied the frequency of DNA markers is that courts insisted on it. When DNA evidence was first introduced, forensic experts offered statistical estimates that were very poorly validated. In the early 1990s a number of appellate courts excluded DNA evidence due to the lack of scientific consensus about the rarity of a DNA match. William C. Thompson, *Evaluating the admissibility of new genetic identification tests: Lessons from the "DNA War."* 84 J.CRIM. LAW & CRIMINOLOGY 22 (1993). Faced with the loss of this favorite weapon in the forensic arsenal, forensic scientists quickly got to work doing the population studies they should have done in the first place. Within a few years, the growing body of research had provided a more solid foundation for claims about the rarity of DNA profiles (at least for the most common types of DNA evidence), and the admissibility crisis quickly faded. Courts have not required comparable statistical validation for other types of forensic iden-

tification evidence and, not surprisingly, little such research has been done.

16. Transcript of hearing at 50, *U.S. v. Jenkins*, No. 96-358-3 (D.S.C. 1997).

17. See Simon A. Cole, William A. Tobin, Lyndsay N. Boggess & Hal S. Stern, *A Retail Sampling Approach to Assess Impact of Geographic Concentrations on Probative Value of Comparative Bullet Lead Analysis*, 4 LAW, PROB. & RISK (2005, forthcoming); see also *U.S. v. Jenkins*, *supra* note 16; *U.S. v. McClure*, Criminal No. DKC 01-0367 (USDC, Dist. of Md.)(Chasanow, J.); *People v. Marlon Smith*, 02CR3477 - Division 9, El Paso County District Court, Colorado; J. Patrick Kelly, J.; *State (NM) v. Trujillo*, Case Nos., D-0101-CR-2000-284; D-0101-CR-2000-229; D-0101-CR-99-677, Judge Candelaria, Santa Fe, New Mexico.

18. See generally David H. Kaye, *From Snowflakes to Fingerprints: A Dubious Courtroom Proof of the Uniqueness of Fingerprints*, 71 INT'L STAT. REV. 521 (2003); Simon Cole, *supra* note 6, at 1226-31.

19. David A. Stoney, *The Measurement of Fingerprint Individuality*, 327, 383 in ADVANCES IN FINGERPRINT TECHNOLOGY (Henry C. Lee & R.E. Gaensslen eds., 2d ed. 2001).

20. Christophe Champod & Ian W. Evett, *A Probabilistic Approach to Fingerprint Evidence*, 51 J. FORENSIC IDENTIFICATION 101, 115 (2001).

The FBI's Budowle, who has emerged as a chief apologist for prosecution junk science, has also been criticized for statistical chicanery in reporting studies designed to validate DNA statistics. See, Dan Krane,

Travis Doom, Laurence Mueller, Michael Raymer, William Shields & William Thompson, *Commentary on: Budowle, et al. CODIS STR loci data from 41 sample populations*, 49 J. FORENSIC SCI (2004); Thompson, W.C. *Additional commentary on Budowle et al.* 43 J. FORENSIC SCI. 447 (1998).

21. Robert D. Koons & JoAnn Buscaglia, *Forensic Significance of Bullet Lead Compositions*, 50 J. FORENSIC SCI. 341 (2005).

22. The study was criticized by the National Research Counsel in its 2004 report on bullet lead. NRC Report, *supra* note 3 and Thompson, *supra* note 3, at 76.

23. *U.S. v. Ronald Mikos*, 2003 WL 22922197, No. 02-CR-137 (N.D. Ill. Dec. 5, 2003) (Guzman, J.).

24. Experts who draw conclusions about ultimate issues often do so on the basis of a logical error known as the prosecutor's fallacy. See, William C. Thompson & Edward L. Schumann, *Interpretation of Statistical Evidence in Criminal Trials*, 11 L & HUM. BEHAV. 167 (1987).

25. The highest quality proficiency testing is "double-blind" administered by an outside agency. In ideal double-blind testing, the test taker and test taker's supervisor(s) should not be aware that the sample(s) under analysis is a proficiency test. Anything less than double-blind testing administered by an outside agency significantly erodes the effectiveness and value of a proficiency test.

26. Quoted in Cole, *supra* note 6, at 1249. ■

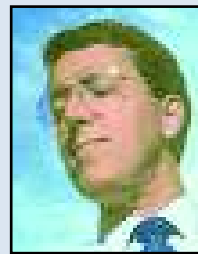
### About the Authors

William A. Tobin retired in 1998 as a Supervisory Special Agent after 27 years of service with the Federal Bureau of Investigation. For 24 years he was a forensic metallurgist at the FBI Laboratory (12 as the *de facto* Chief Forensic Metallurgist). He is presently a forensic metallurgical and materials science consultant in criminal, civil and non-litigious cases.



**William A. Tobin**  
2708 Gunstock Rd.  
Bumpass, VA 23024-8882  
804-448-3955  
Fax 815-331-0654  
E-MAIL wtobin@nexet.net

William C. Thompson, J.D., Ph.D., is professor and chair of the Department of Criminology, Law & Society at the University of California, Irvine and a member of the California bar. He is currently co-chair of NACDL's Forensic Evidence Committee.



**William C. Thompson**  
Department of Criminology,  
Law & Society  
University of California  
Irvine, CA 92697  
949-824-6156  
Fax 949-824-3001  
E-MAIL william.thompson@uci.edu