Absence of Statistical and Scientific Ethos: The Common Denominator in Deficient Forensic Practices

William A. Tobin a, H. David Sheets b, and Clifford Spiegelman c

a Forensic Engineering International, Lane Anna, VA; b Department of Physics, Canisius College, Buffalo, NY; c Department of Statistics, Texas A&M University, College Station, TX

ABSTRACT

Comparative Bullet Lead Analysis (CBLA) was discredited as a forensic discipline largely due to the absence of cross-discipline input, primarily metallurgical and statistical, during development and forensic/judicial application of the practice. Of particular significance to the eventual demise of CBLA practice was ignorance of the role of statistics in assessing probative value of claimed bullet “matches” at both the production and retail distribution levels, leading to overstated testimonial claims by expert witnesses. Bitemark comparisons have come under substantial criticism in the last few years, both due to exonerations based on DNA evidence and to research efforts questioning the claimed uniqueness of bitemarks. The fields of fire and arson investigation and of firearm and toolmark comparison are similar to CBLA and bitemarks in the absence of effective statistical support for these practices. The features of the first two disciplines are examined in systemic detail to enhance understanding as to why they became discredited forensic practices, and to identify aspects of the second two disciplines that pose significant concern to critics.

Introduction

The popularity of CSI-type media through the decades, from Sherlock Holmes to CSI Miami, attests to the public’s persistent fascination with how logical deduction, science, and/or forensic practices can be used to unmask the identity of criminals and solve crimes. Overlooked in these fictional portrayals are the wrongful convictions of defendants sent to prison in part by junk “science” or oversold forensic practice and, conversely, the actual perpetrators of crimes who were not apprehended in a timely manner, if ever, after forensic errors derailed the investigative processes in cases of wrongful convictions. The U.S. National Research Council (NRC) reported in 2009 that, except for nuclear DNA, the scientific and statistical underpinnings for most forensic practices are lacking. This article discusses the characteristics of the discredited forensic practices of comparative bullet lead analyses (CBLA), the recently highly criticized practice of bitemark comparisons, and two additional areas where statistical backgrounds are similarly lacking: fire/arson investigations with limited systematic, empirical evidence, and firearms/toolmarks identification,1 a practice with excellent investigative value but deficient when presented as evidence of guilt.

The impetus for this article was to distinguish common features in the shortcomings of questioned forensic practices. Several of the discredited or questioned disciplines rose to rapid acceptance after high profile cases: CBLA from the John F. Kennedy assassination investigation, and bitemark analysis as a result of its compelling use in one Ted Bundy serial murder trial. Both were built on what deceivingly appeared to be solid bases: CBLA made use of highly sophisticated chemical analyses, while bitemark comparisons were an extension of the effective use of dental records for post-mortem identification. Both lacked substantial and effective accompanying statistical models for the ultimate inferences rendered in forensic application. Fire/arson investigation and firearms/toolmark identification developed from long traditions of anecdotal and empirical observational inference within law enforcement, and have both long been accepted in court. Both practices were developed by insular communities of nonscientist practitioners who did not incorporate effective statistical methods.

Avoiding future errors in forensic evidence interpretation requires that we evaluate past errors with an eye to understanding the practices and limitations that lead to those errors. Assessment of two disciplines that have clearly founndered due to a lack of mainstream scientific and statistical understanding, and of two disciplines currently with similar gaps, can help develop an understanding of how to effectively use and present results from the current forensic practices, as well as others not addressed in this essay or which might be developed in the future.

Characteristics of a Failed Forensic Practice: Comparative Bullet Lead Analysis (CBLA)

The forensic practice of comparative bullet lead analysis, or CBLA, was undoubtedly the most sophisticated but fatally...
flawed practice promoted by forensic practitioners for use in the U.S. judicial system for almost four decades. In simple terms, it compared the chemical composition of crime scene bullets or bullet fragments to bullets found in possession of a suspect under the assumption that “same composition = same source of bullets.” It was used primarily when crime scene bullets were too mutilated for conventional firearm identification comparisons or there was no suspect weapon recovered. The principal assumption that “same composition = same source,” required by logical necessity for foundational validity of the practice, had never been subjected to comprehensive or meaningful validity testing; it had merely been intuited. A second major flaw, eventually fatal to forensic utility for probative value, was that proponents of the practice (FBI Laboratory) had never conducted any research of retail distribution of same-composition bullets in the U.S. Probative value of claimed “matches” had been merely intuited.

CBLA compared between 3 and 7 elemental concentrations among the crime scene and suspect bullets using very sophisticated analytical instrumentation. If the measured elements were sufficiently close quantitatively between or among compared bullets, the forensic examiner would testify in judicial proceedings that the crime scene bullets originated from the same “source” of bullets, often claiming the “source” as same box of bullets as seized from a suspect. The NRC evaluated the practice and, in their 2004 report “Weighing Bullet Lead Analysis,” articulated some of the reasons why such testimony was unfounded. After challenges by scientific critics, and a 2007 episode of CBS’ popular series “60 Minutes” and Washington Post collaborative exposé, the FBI discontinued CBLA practice. The FBI eventually disclosed by letter to a small subset, as it turned out, of attorneys for defendants convicted in whole or in part on CBLA testimony, that such testimony was not scientifically supported.

In a forensic practice that should have consisted of four component phases, actual CBLA practice at the FBI Laboratory, virtually the sole proponent of the forensic practice in the U.S., consisted of only three. The first phase (“analytical”) entailed precise quantitative analysis of chemical analytes in the lead matrices of bullets from crime scenes and/or from suspects. The chemical elements were either measured using a nuclear reactor in a procedure known as neutron activation analysis (NAA), or by inductively coupled plasma–atomic (or optical) emission spectroscopy (ICP-OES herein), where small samples of the bullets were dissolved in acid and converted to high-temperature plasma. Consequent light intensity and wavelengths emitted during return to ambient stability (ground state) from the sample plasma were measured. Bullet composition is difficult to measure; however, the FBI Laboratory became very proficient at such analyses and could measure the seven target elements antimony, copper, silver, cadmium, antimony, and arsenic, and tin to the parts per million with impressive precision. Both analytical processes (NAA and ICP-OES) were very sophisticated.

Once the compositional elements were quantified, the next phase of forensic practice (“grouping”) was to determine which bullet data were considered “analytically indistinguishable” (bullets that purportedly “matched” in composition) and which were “analytically distinguishable” (bullets that did not compositionally “match”). Customarily, from the period of approximately 1970 to June 1998, a very loosely applied 1-sigma “match” criterion with data chaining was used to declare samples “analytically indistinguishable,” but reviews of several hundred transcripts revealed that examiners frequently overrode or ignored the criterion in court testimonies if they subjectively “felt” that the samples should be considered a “match.” For several decades, the FBI match criterion was not memorialized in protocol. In June 1998, the FBI Laboratory was compelled to formalize a match criterion, and consequently arbitrarily established a 2-sigma match criterion, virtually doubling the range of possibilities for incriminating compositional “matches,” but continued to use the objectionable (for forensic application) statistical process known as data chaining, and continued to override the 2-sigma criterion if they “felt” that the bullets should “match.”

While the data treatment and match criteria used were generally problematic, chaining was the most specifically objectionable. To illustrate chaining, suppose that we have four bullets labeled A through D. Further suppose that the antimony levels of bullets A and D do not match because the mean antimony levels are farther than 4-sigma apart (2-sigma for each sample). Submitted to the FBI crime laboratory alone, these bullets would not be considered incriminating as a “match.” Now hypothetically assume that the antimony levels in A and B are within 4-sigma, that the levels in B and C are within 4-sigma, and finally that the levels in C and D are within 4-sigma. Because A could be chained to B, B could be chained to C, and C could be chained to D, then by the FBI matching criterion, examiners would conclude and testify that bullets A and D “matched” (as analytically indistinguishable) on antimony. In essence, the very same two samples that, alone, would not be considered incriminating, became incriminating only because more bullets were submitted that could be “chained” between A and D. By that reasoning, assuming no metallurgically highly unlikely extreme clusters in chemical composition distribution, every bullet ever manufactured in the history of mankind could be considered “analytically indistinguishable” with any other bullet ever made if sufficient bullets were submitted for comparison. Similarly, if all seven analytes (chemical elements measured) “matched,” then the FBI would testify that the crime scene bullets “matched” those in, or likely came from, the suspect’s box, regardless of how many bullets were needed to make the chain.

Underlying the claim that the bullets likely came from the defendant’s box were some FBI Laboratory assumptions. Those were:

1. Representativeness: that the minute (12–60 mg) forensic samples were compositionally representative of their 100+ ton molten sources;
2. Homogeneity: that all molten 100+ton batches of lead were chemically uniform (homogenous);
3. Uniqueness: that, due to the high precision measurements, no two production batches (called “pots,” “heats,” or “melts” in metallurgical and industry terminology) had overlapping or repetitive chemical concentrations.
4. Probative value: that bullets were distributed in retail markets in such a way that made it unlikely that a chemical concentration of bullets was common in any geographic area.

The third phase of actual forensic practice was the inference phase. After conducting the compositional analyses and evaluating the resulting data for apparent “matches,” the CBLA
examiner would proceed to offer incriminating testimony that defendants’ bullets were “analytically indistinguishable” from crime scene bullets and, therefore, “originated from the same source of bullets manufactured on or about the same day” or, eventually and frequently, “originated from the same box of bullets” or “consistent with originating from the same box” of bullets.

The required fourth phase for judicial application should have been assessment of probative value of claimed “matches.” This critical, but missing from FBI Laboratory practice, phase was eventually exposed by critics who, in efforts to assess probative value of claimed bullet composition “matches,” researched the geographical distribution of bullets in retail markets (Cole et al. 2005). If everyone who owned bullets in the region of a shooting had bullets of the same composition, there would be no probative value of a claimed compositional “match.” Researchers wanted to determine if inhomogeneities existed in retail bullet distribution or, conversely, if regional concentrations of same-composition bullets existed. Even the researchers were astonished at the extent of same-composition product saturation of local markets. FBI executives eventually recognized the lack of forensic utility based on the critical researchers’ findings, and announced on September 1, 2005, that they would no longer offer CBLA as a forensic service because no one could assess the probative value of a claimed “match” in any particular case (FBI 2005).

The FBI had never meaningfully researched or tested the scientific validity of any of the four underlying assumptions required by logical necessity for scientific validity of the forensic practice (representativeness, homogeneity, uniqueness, probative value). Falsifying any one of the assumptions would have invalidated the entire forensic practice of CBLA. It is axiomatic that all four should have been subjected to rigorous validation studies and experiments prior to offering the service to the law enforcement and judicial communities. A retired FBI forensic metallurgist and colleagues challenged the practice because the assumptions of homogeneity and representativeness defied long-established scientific (metallurgical) principles, and conducted research into both bullet lead secondary refining operations and retail bullet distribution in order to assess probative value of claimed “matches” (Tobin and Duerfeldt 2002, p. 26). The research was revelatory: it falsified not just one of the critical assumptions, it falsified all four as universal assumptions. Worse, the researchers’ geographic studies of retail bullet distribution demonstrated that CBLA had no demonstrative probative value at all and, thus, the compositional data were meaningless for forensic application (Cole et al. 2005).

As for why the CBLA “junk science” was admitted for almost four decades in judicial proceedings, appearances in many evidentiary hearings and trials, and reviews of several hundred testimony transcripts, revealed that judges and forensic practitioners alike had been seduced by the sophistication and precision of the analytical instrumentation used in phase 1 (analysis) of the forensic practice into believing that such precision surely must equate to probative value, yet the FBI Laboratory never bothered to research implications or correlations of analytical precision with either the underlying assumptions of the forensic practice or with product distribution. The analytical phase of CBLA had not been seriously challenged by critics because it was not the “task at hand” as envisioned by the Daubert opinion. In the end, costly sophisticated analytical scientific instrumentation, including a nuclear reactor to generate compositional data to the parts per million, produced remarkably precise data that were meaningless for forensic application.

Evidence of Difficulties with Bitemark Comparisons

At this time, the use of bitemark analysis, specifically comparison to identify an individual as producing a particular bitemark resulting from an assault, has come under increasing scrutiny and criticism. The much-discussed report on forensics from the National Academy of Sciences (National Research Council 2009) observed that bitemark comparison had never been exposed to stringent scientific scrutiny. The NAS report noted that the primary rationale for the validity of these analyses rested on the history of their use in court, as opposed to tests of the validity of the methods. Under such conditions, in which there is little empirical evidence of the validity of the methods, the numerous false convictions based on such comparisons are a serious concern, and imply methodological error.

Of the 337 exonerations evaluated by the Innocence Project (Innocence Project 2016), 17 involved bitemark analysis. Turvey and Colley (2014) discussed a total of 26 cases in which bitemark analyses allegedly led to improper convictions, and the exoneration of Keith Allen Harward after 33 years of imprisonment based on bitemark evidence was announced during the course of writing this manuscript (Hsu 2016). These exonerations are particularly disturbing given that the logic behind bitemark analysis is largely inferential in nature, and there is no known, currently available estimate of the rate of false positive matches within the field. The existence of valid exonerations indicates a nontrivial rate of error in these methods.

Jo Handelsman, Assistant Director of the White House Office of Science and Technology Policy was quoted in the Washington Post (Balke 2015) as publicly calling for the eradication of poorly validated forensic methods:

“Suggesting that bite marks [should] still be a seriously used technology is not based on science, on measurement, on something that has standards, but more of a gut-level reaction. Those are the kinds of methods that have to be eradicated from forensic science, and replaced with those that come directly out of science, and have the ability to stand up to the standards of scientific evaluation.”

The Texas Forensic Commission is currently investigating the status of bitemark comparisons in the state of Texas, motivated both by the efforts of the Innocence project and by recent exonerations of convicts in Texas. Testimony by Ian Pretty and Adam Freeman to the Texas Forensic Commission (as reported in The Dallas Morning News, Grissom 2015) on a research study they conducted jointly with the American Board of Forensic Odontologists (on the ability of forensic dentists to identify impressions as bitemark evidence) reportedly showed disturbingly low rates of agreement on which images were really bitemarks. The complete results of the survey have not been released, apparently on the grounds that the instructions in the survey used were unclear.

2 Imwinkelried and Tobin (2003), at 61, discussing Daubert v. Merrill Dow Pharmaceuticals, 509 U.S. 579, 590 (1993).
It seems clear from the exonerations that there have been substantial overstatements of the strength of bitemark evidence at the very least. It is reasonable to ask how this came about. How was it possible to overstate the strength of the evidence to the point of claiming that bitemarks provided a unique identification of an individual as being the producer of a given bitemark? One approach is to look at the basic claims of the field:

I. Human dentition is unique

II. A bitemark in human skin accurately records enough information about the dentition to link an individual to the bitemark.

Therefore, if an individual dentition can be said to match a bitemark, then the individual must have created the bitemark. As noted by the NAS report, the continued acceptance of this argument by the courts perpetuates continued use of bitemark analysis by proponents.

**Inferential Arguments**

Argument for validity of bitemark analysis is syllogistically founded. Two premises are stated, which seem reasonable enough, and are presumed to be true, and then a series of statements are deduced as logical consequences of those initial premises. This sort of inferential logic, deductions based on premises assumed to be true, coupled with the argument that the court system has accepted these arguments as valid, appears to be the foundation of much thinking about forensics, particularly in areas such as toolmarks and bitemarks where little empirical evidence is available.

In contrast, much of modern science focuses on hypothetical-deductive approaches to knowledge, using a process of the form of forming and stating an initial conjecture (hypothesis) in falsifiable form, deciding what the implications of the hypothesis are, assuming it is true, and if it is false, and then collecting data or performing one or more experiments to attempt to refute the hypothesis. The conjecture or hypothesis is assumed true only to deduce the consequences and form the test.

The American Board of Forensic Odontologists (ABFO) certifies some member forensic dentists as Diplomates, giving them some level of credentials as forensic dental specialists. The forensic areas covered by the ABFO include post-mortem identification from dental remains, testimony in dental malpractice, age-estimation from dental structures, and bite-mark analysis. The post-mortem identification of individuals from dental records has a long and effective history, and is a major contribution of the forensic dental community to the larger society. In a post-mortem identification, the forensic dentist has access to all teeth available in the dentition, and can examine all sides of the teeth, often including the root morphology if pre-mortem X-rays are available, and can often see all the restorations and dental interventions performed. When good quality, detailed dental records are available, it is possible to document and compare all of the information obtained post-mortem to the treatment history. This type of identification procedure can also be checked in many cases by a slower and more costly DNA identification, so that there is a feedback loop available in post-mortem identification that is typically not available in other forensic identifications.

The undeniable success of post-mortem identification, also based on the premise that all individuals have distinct dentitions, presumably lent some credence to the idea that individuality of the dentition held true for bitemarks. However, in a bitemark, only the incisal (cutting) surfaces of the six most anterior teeth (the canines, the lateral incisors and the central incisors, assuming all are present) of each jaw have the potential to leave an impression in a typical human bitemark. This means that only 1 surface of each of 12 teeth are available for analysis, relative to the possibility of examining as many as 5 or 6 surfaces of all 32 teeth, plus restorations, as is available in a post-mortem identification. This is not to say that all of this information is necessarily available in each instance of post-mortem identification; in some cases much less information is available. It seems highly probable that the effectiveness of post-mortem identification strengthened the acceptance of the premise of individuality of human dentitions as used in bitemark comparisons.

**Empirical Evidence**

In addition to this sort of syllogistic argument from common sense that bitemark analysis should work as claimed, there were several papers in the literature purportedly supporting this premise, some of which are at least partially inferential in nature rather than empirical. The recent review paper by Franco et al. (2014), conducted a highly detailed literature search looking for studies that examined the claim that the human dentition as recorded in a bitemark was unique, and found a total of 13 published papers that addressed the issue using datasets ranging in size from 11 to 1099 specimens (Disclosure: One of the co-authors of this article, H. D. Sheets, was a co-author on a number of these papers). Of the four papers that attempted to claim the biting dentition was unique, the sample sizes were 10, 397, 50, and 13. Franco et al. concluded that uniqueness had not been established by these four papers and that there were constraints in each of the four that limited their interpretation.

The 1984 paper by Rawson et al., “Statistical Evidence for the Individuality of the Human Dentition,” directly addressed the issue and attempted to use a database of measured bitemarks to arrive at a statistical argument that the human anterior dentition as recorded in a bitemark was effectively unique. This paper has the largest sample size of the four papers claiming some level of empirical evidence that the biting evidence is unique, and is probably the most commonly cited such evidence. The authors of the study collected bitemarks impressed in wax from 397 individuals, and recorded the position of the center of each of the 12 teeth involved and the angular orientation of each tooth, all from the bitemark. These measurements were then mathematically superimposed on a baseline and centered on the edge of one incisor to eliminate differences due to translation and rotation, yielding three measurements for each of 12 teeth. The range of each measurement was then divided up into discrete positions (which might also be called cells, bins, or states), such that the bin width was the measurement resolution of the system. The use of bins of measurements is common and analogous to the bins used for allele mass calculation used in DNA analysis, so this approach had some level of precedent.

After determining the number of possible positions into which each the three measured variables for each tooth could
fall, the total number of possible combinations of state values was calculated for all the teeth in the upper (maxillary) and lower (mandibular) dentitions. The values obtained (2.2 × 10^{13} for the maxillary and 6.08 × 10^{12} for the mandible, 1.36 × 10^{26} for both combined) are far beyond the population of the earth and, thus, the authors concluded that the anterior dentition as recorded in bitemarks is effectively unique.

It can be argued that the Rawson et al. (1984) study is inductively inferential in nature. While a data collection process is used to generate empirical data about the number of states each tooth position could in principle occupy, there is no hypothesis testing used to see if the derived theory of the paper actually describes that dataset. The measured number of positions of each tooth is used only as an input to the inductively inferential process. A total of 397 specimens are used in the study, only summary statistics based on the data are used, since the range of each measured value and the measured uncertainty in position are used to calculate the number of positions. While some statistical concepts were applied to form the argument presented in the paper, there was no empirical hypothesis test in the study.

In the derivation and interpretation of the number of states found by Rawson et al. and the interpretation of that large number of states, it is assumed in the calculations (but not explicitly stated in the paper) that all state positions are equally likely to appear and that the probability of occupying a state is independent of the states occupied by other teeth in the same dentition. The statistical model used is independent uniform distributions for each measurement. Based on the measured data and these assumptions, the authors then infer that, since the number of state combinations is immense, the probability of any two individuals matching is so small as to be meaningless.

In the replication of this study by Bush, Bush, and Sheets (2011), the model developed by Rawson et al. (1984) was treated as a hypothesis rather than a premise. While the probability of all 12 teeth in the biting dentition predicted was too small a value to be testable in any reasonable size database, the Rawson et al. model could also be used to predict the probability that the three position measurements of a particular tooth would match (occupy the same state) for two different unrelated individuals, or the probability that two specified teeth would likewise occupy the same state. Tooth 22 was predicted to have a total of 107.2 possible states, based on Rawson et al.’s data, and 154.5 possible states using the Bush, Bush, and Sheets data (the dataset used by Bush, Bush, and Sheets actually showed a wider range of variation than that in the Rawson et al. study, and hence estimated even more states). This would imply probabilities of 1/107.2 or 1/154.2 that unrelated individuals would have matching positions of tooth 22. Bush, Bush, and Sheets used a dataset of 172 specimens, allowing for 14,706 pairwise comparisons of individuals. Based on Rawson et al.’s state estimation the expected number of matches would be 14,706/107.2 = 137.1, or using the Bush, Bush, and Sheets state estimate, 14,706/154.2 = 95.1 matches. But when the data were actually checked, 2453 of the pairwise comparisons were matches. The difficulty here was the assumption that tooth position followed a uniform distribution model; most people are average, and the distributions are closer to normal than uniform. The probability of occupation of each particular state is not simply 1 over the number of states; more sophisticated distributional models are required.

When the position of more than 1 tooth at a time was examined in the same manner, by computing the expected number of matches of 2, 3, 4, 5 and 6 teeth in the lower dentition and then comparing these expectations of match rates with the 172 specimen database, the number of matches far exceeded the model expectation. A total of 7 pairs of specimens had matching states on all six teeth in the lower dentition. A bit of consideration of the biology of the data explains why. Human dentition has to form a working bite, and humans display some level of bilateral symmetry, so that there is substantial amount of correlation of the positions of the teeth with one another; they are not independent. If the right side of a human face is unusually broad, the expectation will be that the left side is as well. The model used by Rawson et al. did not reflect the presence of this correlation and they did not examine their data for the presence of correlation.

It appears that the rationale for the use of bitemark comparisons within forensics is that there are a pair of untested premises about bitemarks, which, based on “common sense,” are logically valid, and that based on these arguments, bitemark comparisons have been admitted in courts many times. It appears that this class of impression evidence is thus supported only by inductively inferential logic and prior accepted use, quite literally tradition. The claimed exonerations of roughly 24 convicted individuals argue that there is some error rate in the use of bitemark evidence. Although there is error in all forms of human endeavor, there exists sufficient empirical evidence that the rate of error for forensic bitemark comparisons exceeds what might be considered acceptable limits for forensic application. The premise that dentition is effectively unique seems well supported by the long and effective history of post-mortem forensic identification based on entire (complete) dentitions, but bitemarks record far less information about the dentition than is typically available in a whole-dentition study, so it appears that the premise of the uniqueness of biting dentition is not a necessary corollary of the effective uniqueness of the entire dentition. It also appears that the preponderance of empirical evidence finds against uniqueness.

The conclusion here would be that the strong reliance on inductively inferential logic has not served us well in the case of bitemark comparison. Flaws in the initial premises of inductive inference may be difficult to ascertain by discussion and logic alone. Effective use of hypothesis-driven empirical research is desperately needed here and in others areas of forensic evidence supported only by inductive inference and judicial history.

Fire and Arson Investigations: Observational Claims Borne from Frustration

The inherent characteristic of arson to consume or destroy the very evidence needed to establish existence of the crime renders it one of the most difficult to investigate and prosecute (Tobin and Monson 1989). Understandably, no forensic practice is more imbued with underlying premises derived from observational experience than fire/arson determinations. But mainstream scientists have long been aware that inductive inference from limited observational studies are quite vulnerable to bias and logical fallacies. Historic development and current practice of fire/arson investigation share some of the common flaws as CBLA, firearms/toolmarks (aka, “ballistics”) identification, and
other forensic practices, to be discussed in the overview summary, below.

Fire scenes are very dirty, and fire investigation is not glamorous. In as much as virtually all fire and arson investigators derive from the ranks of law enforcement, few are schooled in any of the hard sciences (physics, chemistry, materials science, metallurgy, mechanical engineering, statistics, inter alia), and fewer yet are schooled in design of experiments, hypothesis testing, or inferential logic processes. Because no rigorous academic programs existed for fire investigation incorporating established scientific principles, for decades, practitioners were required to develop their own “burn indicators” by extrapolating anecdotal observations. However, anecdotal observations and inductive extrapolations from various fires to future events and pronouncements are vulnerable to bias, logical fallacies of presumption, and other flawed inferential logic, many of which were passed down within the practice culture and became imbued in the domain literature. Unless such anecdotal observations used for extrapolation are sufficiently similar in all relevant aspects (influential input parameters) to those of the event(s) to which they are extrapolated, their use as burn indicators can lead to flawed inference. Largely because there is little extrajudicial interest in the domain literature by other than fire investigators themselves, there was virtually no oversight or challenge from the true, mainstream scientific community. Although many of the observationally developed guidelines have been relatively recently exposed as mythology and folklore, some continue to be used today.

An example of inductive extrapolation without scientific oversight is illustrative. During the investigation of the July 1996 TWA Flight 800 midair explosion over Long Island, N.Y., one coauthor continually battled flawed logic from anecdotally derived fallacious extrapolation by numerous bomb techs who had investigated the midair explosion of Pan Am Flight 103 over Lockerbie, Scotland, years earlier. At the reconstruction hangar on Long Island for the TWA Flight 800 investigation, bomb tech investigators periodically presented fragments from the fatal flight that were recovered from the bottom of the Atlantic Ocean, for metallurgical examination, excitedly claiming that they had found “the smoking gun” … pieces of monocoque (aircraft structural skin) from the Boeing 747–100 exhibiting what is known as “orange peel,” a phenomenon created from impulsive loading (forces from extremely high strain rate events such as from a bomb). The problem was one of flawed logic: the converse does not have the same truth value as the original assertion: because impulsive loading can create “orange peel,” observing “orange peel” does not necessarily imply impulsive loading. In other words, the argument that all canaries are birds does not imply that all birds are canaries. For decades, the same observationally extrapolation process from fire scene investigations was used to develop guidelines for fire investigators desperate for criteria by which to accurately assess cause and origin. In that process, hypothetically, a certain characteristic might be observed from a specific fire admittedly set by a suspect. The characteristic may not be relevant to assessing cause and origin of our hypothetical fire, but it could easily be used, subconsciously or otherwise, as an “indicator” of arson or accelerant use during investigations of subsequent fires. "Because I saw X at fires known to have been set, then whenever I see X, I know a fire was incendiary in nature.”

One such phenomenon used for several decades to indicate cause of fires was post-fire observation of whether or not furniture springs, such as found in mattresses and couches, were found to be “collapsed” or not. Research of the “authoritative” fire and arson literature on the subject raised some red flags, not the least of which was that there were two, contradictory, schools of thought represented to explain ostensibly “collapsed” springs observed pursuant to residential fires. Some authors claimed that, because of the lengthy “annealing” time required to soften the springs that would allow them to “collapse of their own weight,” collapsed springs indicated a long, slow smoldering fire such as from a smoldering cigarette in a couch. Other authors claimed that because of the elevated temperatures required to anneal furniture springs, an accelerant would have been required (for our discussion here, ignore the fallacy that temperature and heat were considered the same metric in the fire/arson investigation domain literature). The basis for one of the red flags raised from literature review was the fact that metallurgists have known for centuries that time and temperature are strongly (inversely) correlated, a fact used in production heat treatment processes, to obtain the same dependent variable microstructural outcome. In simplistic terms, a similar microstructural outcome might be obtained by holding a metal at 2000 degrees for 1 hr as holding a metal at 1000 degrees for 2 hr.

In efforts to reconcile the disparate rationales, metallurgical researchers conducted burn experiments of mattresses at the FBI Burn Facility in Quantico, Virginia, simulating both smoldering cigarettes and accelerant-induced fires (Tobin and Monson 1989). As suspected by the researchers, it turned out that there was no observable difference in spring collapse behavior. Additionally, furniture springs were subjected to experimental observations of percentage collapse at times up to 65 min and temperatures to 1700 degrees Fahrenheit with and without impressed loads (simulating victims or collapsed ceiling structure). The experiments confirmed that there was no probative value even in observations of “degree of spring collapse” as many fire-investigator authors had advocated in the domain literature for several decades. The belief comprising pathological science turned out to be a classic case of anecdotal extrapolation based on fallacies of presumption (primarily false dichotomy and suppression of evidence).3

---

3 Which can, counterintuitively, attain 900–1100 degrees Fahrenheit. See, for example, NIST Special Publication 852 (August 1993), U.S. Department of Commerce, at 49.

4 Pathological science is a term believed to have been first used by Nobel Laureate Irving Langmuir in his presentation at a colloquium at The Knolls Research Laboratory, December 18, 1953. It characterizes situations, with no implications of dishonesty, where people are influenced into mistaken beliefs or false interpretations of results resulting from a lack of understanding about how humans can deceive themselves and be led astray by subjective influences, wishful thinking, cognitive biases, or unforeseen interactions between or among input variables known as “threshold interactions.” Some of the beliefs have attracted a great deal of attention, such as the claims for cold fusion and polywater, with hundreds of papers published on the topics, and have lasted decades only to eventually fade from public memory as the beliefs and interpretations are revealed to be invalid.

5 “False dichotomy” is also known as the “either-or; “false choice;’ "false dilemma" fallacy, inter alia, where options are presented as nonjointly exhaustive alternatives when, in fact, alternative options are available and reasonable. "Suppression
Firearm/Toolmarks: Long in Tradition, Short in Validation

Firearms identification, aka “ballistics,” a subset of toolmarks identification, has been admitted for over 70 years in the courtroom as a presumably “reliable” forensic practice. However, relatively recent challenges from the true (mainstream) scientific community question the scientific foundations of the practice, but defenders of the practice cite its lengthy judicial admissibility as evidence of its scientific reliability for Daubert purposes.6

From a scientific perspective, a salient threshold consideration for this discussion is that practitioners and jurists alike conflate validity and reliability. The courtroom is not a laboratory; a fact lost on virtually all firearms/toolmarks practitioners and most judges when it comes to the difficult task of evaluating the scientific “reliability” of the quasi-technical testimonies. Lengthy judicial admissibility does not equate to scientific validity or reliability. Unfortunately, many judicial rulings incorporate that fallacy of presumption, with the presumption that decades of consistent legal opinion equate to scientific reliability.7 As observed by Faigman et al.:

While expert evidence on toolmarks and firearms identification is universally admissible, this universal admissibility has come about with virtually no judicial evaluation of the validity of the underlying science or its application. One might have expected the situation to change following Daubert, but so far that has not happened. (Thornton and Peterson, 2002a, p. 194)

In September 2016, the President’s Council of Advisors on Science and Technology (PCAST; 2016), issued a report concerning. In its report, the PCAST observed that foundational validity of the field of firearms/toolmarks identification has not been established, echoing similar observations by the National Research Council (NRC) of the National Academies of Science (NAS) in its report 7 years earlier, “Strengthening Forensic Science in the United States: A Path Forward” (President’s Council of Advisors on Science and Technology 2016, p.11). In fact, because the foundational validity and reliability of the forensic practice has not been established and currently has no probative value, the PCAST recommended that the practice of firearms identification not be offered in courts. (President’s Council of Advisors on Science and Technology 2016, p. 19).

Reliability relates to consistency in observational results (repeatability and reproducibility), not the accuracy or validity of the results. As noted by two authors differentiating reliability and validity,

... a parade of forensic scientists who make the same subjective judgment, or a series of machines that give the same readings in response to the same evidence sample, can only be said to be reliable. Validity refers to the degree to which a measuring instrument measures what it purports to measure. Id. See also Horn, 155 F.Supp.2d at 538–539. In short, reliability refers to consistency, while validity refers to accuracy or correctness. Thus, “forensic scientists or machines that are in agreement may be highly reliable (in agreement with each other) without being valid (without reaching the correct answer). They can all be wrong.” Id. This is a critical concept to grasp, as the forensic identification fields are notorious for trying to equate consistency with correctness. (Thornton and Peterson, 2002, pp. 16–7)

In other words, an instrument may be calibrated incorrectly, for example, a weight scale that is inaccurate by +17 lbs., but still be reliable in indicating that the readout is consistently 117 lbs. for a 100-lb. weight.

The accuracy of firearms identification is possibly more strongly correlated to efficacy of investigative function by law enforcement than to forensic practice, resulting in likely Type III errors (attaining a correct answer but for the wrong reason, in essence, inadvertently obtaining a correct answer). Although both the judicial and forensic domains are comprised primarily of nonscientists, it is quite ironic that some insightful jurists, generally even more removed from science and technical endeavor than forensic practitioners, are beginning to recognize the vacuous claims infused in firearms identification practice with unfounded expressions of certainty, yet practitioners still have not.

Firearms/toolmarks identification is essentially a pattern-matching practice that is virtually entirely subjective once the possible sample pool is reduced by class characteristic screening. It is not a science,8 given that every critical cornerstone of the scientific method is either missing from the forensic practice (e.g., no scientifically acceptable falsifiable hypothesis) or is unreliable in practice (e.g., repeatability and reproducibility). Examiners typically use a comparison microscope to view characteristics (striations and/or impressions) of tribological interaction of bullets with barrels and/or cartridge cases with firearm components, find some characteristics that are in “sufficient agreement,” and opine specific source attributions (aka, individual

6 For example, see U.S. v. Green, 405 F. Supp.2d 104 (D. Mass. 2005)(Gertner, J.) which discusses ways in which ballistics evidence fails to meet Daubert criteria regarding, inter alia, testability, reliability, and error rates. See, U.S. v. Chaz Glynn, 578 F. Supp. 2d 567 (D.Mass., 2008); 60 Cr 580 (JSR) (SDNY 2004) (Rakoff, J.) (Sept. 22, 2008), where the Court...very quickly concluded that whatever else ballistics identification analysis could be called, it could not fairly be called “science”, at 6. See also, Williams v. U.S., 13-13F-1312 (D.C. Cir. Jan. 21, 2016) (Easterly, J., concurring), where Judge Easterly observed that any expression of certainty asserted by a firearms examiner in testimonies relating to their pattern-matching practice “...has the same probative value as the vision of a psychic”, at 26, available at http://www.dccourts.gov/internet/documents/13-13F-1312.pdf. See also, State v. Goodwin-Bey, No. 1531-CR0055S-01, Greene County Circuit, MO (Dec. 16, 2016) (Holden, J.), where the Court ruled that the practice is “all subjective”, not scientific, without foundation for error rate, and reluctantly allowed the witness to testify only that the firearm could not be eliminated as the firing platform for the fatal bullet.

7 Both legal and scientific scholars have expressed agreement on this observation. For example, see Fn. 10, supra, and the National Commission on Forensic Science (NCFS) position statement wherein it discourages use of the term “scientific” in expressions of certainty because it “implies that the discipline is indeed a science,” at 4, available at https://www.justice.gov/ncfs/file/64133/download.
Bullets or cartridge cases are consequently concluded to have been “fired from/in” a specific firearm.

The only “protocol” for the practice is merely a guideline from the trade association for the domain and is known as the “AFTE Theory of Identification.” It is not an acceptable scientific protocol for a number of reasons, but the principal deficiency is that it does not present a falsifiable hypothesis. The guideline is circular (tautological), and is so vague that, once a target sample pool is narrowed by class characteristics (e.g., caliber), it basically allows opinions based on no objective criteria at all, but rather on “training and experience,” a 100% subjective criterion. As forensic scholars have articulated, because of its virtually total subjectivity,

[firesarms/toolmarks] experts exploit situations where intuition or mere suspicion can be voiced under the guise of experience. When an expert testifies to an opinion, and bases that opinion on “years of experience,” the practical result is that the witness is immunized against effective cross-examination. When the witness testifies that “I have never seen another similar instance in my 26 years of experience,” no real scrutiny of the opinion is possible. No practical means exist for the questioner to delve into the extent and quality of that experience. Many witnesses have learned to invoke experience as a means of circumventing the responsibility of supporting an opinion with hard facts. For the witness, it eases cross-examinations. But it also removes the scientific basis for the opinion.

Testimony of this sort distances the witness from science and the scientific method. And if science is removed from the witness, then that witness has no legitimate role to play in the courtroom, and no business being there. If there is no science, there can be no forensic science. [emphasis in original] (Thornton and Peterson 2002, pp. 16–17)

The same argument applies to typical testimonies of firearms/toolmarks examiners when they decline to produce photographs or to describe to the court or jurors exactly what characteristic(s) they used to claim an “identification.” The standard mantra, as has also been posted on the AFTE website as a guide for examiners to defend their testimonies in cross-examination, is that photographs are “misleading” in presenting only two-dimensional perspective rather than the alleged three-dimensional perspective that they claim, in most testimonies observed or reviewed by the authors, to see under a microscope. Notwithstanding that, in the decades-long experience of one author, the representation is grossly misleading, and comprises an additional aspect of the immunization process described by Thornton and Peterson. Individually or severally, the circumstances serve to insulate the examiner from having to defend the implied claim that “I know it when I see it,” rendering it virtually impossible to challenge any claim of an identification.

The underlying premises required by logical necessity for firearms/toolmarks identification forensic practice are that (1) each firearm is unique, and (2) that characteristic persistence (“repeatability” in firearm identification terminology, but not in the true scientific sense) is not volatile over significantly long periods of time. There exist numerous flaws in firearm identification practice, primarily failings of presumption and grossly deficient (or, more prevalently, completely lacking) designs of experiments. Critics have characterized the practice as “pathological science” and some insightful judges are in agreement, resulting in embryonic stages of a judicial paradigm shift (Tobin and Blau 2013).

A most critical underlying assumption for testimonial opinions of individualization is the purported ability of a firearms identification examiner to eliminate from consideration what are known as “subclass” characteristics, those deriving from the manufacturing process, in other words, striations or impressions remaining from the last forming operation in production, and which would exist on all components in a potentially very large production lot. It is known within the firearms identification community that an examiner must eliminate “subclass carryover” before making an identification but, unfortunately, the general practice is to assume a condition of “all-or-nothing,” that only purportedly “individual” characteristics transferred to the final product. The only bases by which examiners can realistically claim to be aware of the critical subclass characteristics for a particular manufacturing process and particular product are ostensibly “tours” of the relevant firearm manufacturing facilities. At an observer level, firearm/toolmark identification practitioners frequently understand the general manufacturing processes for firearms. However, the characteristics used by practitioners for comparison purposes in firearms identification are generated by a variety of metallurgical processes and entail complex tribological and microstructural (including atomic) interactions that can, and most often do, vary from product to product, and even from production lot to production lot, virtually none of which can be observed and interpreted from plant “tours” by nonmetallurgists. The most relevant true scientific domain, metallurgy/materials science, that understands the tribology of the manufacturing processes and their specific seminal effects on firearm components leading to the very characteristics used by forensic practitioners for their identifications, is not engaged. Thus, there are known unknowns by the general scientific community, and unknown unknowns within the toolmark community. These “known unknowns” comprise critical, but currently missing (“unknown unknowns”), influential parameters for any purported “validation studies” in the domain.

The same holds true for statistics. Neither the toolmark community nor the judiciary in criminal matters is proficient in application, nor do they generally even acknowledge the significance, of statistical tools and concepts as the forensic practice demands. The overwhelming majority of practitioners deny that statistics has any role whatsoever in firearms identification, and continue to deny the inherently probabilistic...
nature of firearms identification "matches." But denying or 
"ignoring the statistical bases for empirical statements, how-
ever, does not make them any less probabilistic … The empirical 
uncertainties of factual statements are as important as the 
statements themselves and should be part of the legal cal-
culus." The firearms identification domain is almost entirely 
enunengaged with the statistical community. Unusual exceptions 
are very recent collaborations at NIST among toolmark experts 
and statisticians. However, given the intransigent defense of 
firearms identification dogma because it has been “accepted” 
in the courts for many decades, implying scientific valida-
tion, and that the misguided beliefs have been memorialized 
in practice domain literature and training for decades, there 
exists a strong cultural bias against, and inertial resistance to, 
challenging existing norms. Regardless of the credentials and 
criticisms of mainstream scientific researchers, to include the 
nation’s most prestigious voice of the scientific community 
(National Academy of Sciences and the President’s Council of 
Advisors on Science and Technology), practitioners persist in 
advocating ingrained beliefs in judicial proceedings even today. 
Largely because of an absence of scientific culture in the forensic 
domain (other than nuclear DNA), practitioners continue to 
rationalize that the scores of purported “validation studies” over 
the decades ‘proves’ the scientific trustworthiness (‘reliability’ 
in legal concept) of the forensic practice. However, mainstream 
scientists researching the domain literature have made several 
surprising observations: there has been no comprehensive or 
meaningful scientific validation of the underlying premises 
required by logical necessity for scientific foundation, and that 
the purported “validation studies” typically proffered to courts 
are seriously flawed, have no external validity, and are largely 
irrelevant to any particular case under judicial consideration. 
In essence, as has been observed by scholars at the intersection 
of science and the law, the premises underlying the forensic 
practice have been adopted purely by faith (intuited), and 
the purported “validation” studies offered to courts have been 
rationalized as “proof” of practice efficacy by wishful thinking.

Common Denominators of Discredited or Challenged 
Forensic Practices

On *prima facie* observation, the diverse natures of the forensic 
practices discussed, CBLA, bitemark comparisons, fire/arsen 
investigations, and firearms/toolmarks identification, seem to 
belie the commonality of systemic influences allowing them to 
become entrenched as pathological “sciences.” Analyses of the 
forensic practices in the context of judicial application reveal a 
confluence of influential circumstances that assuredly facilitated 
their entrenchment in decades of legal admissibility.

At an operational level, misconceptions of proficiency testing 
and purported “validation” studies facilitated perpetuation 
of the flawed practices of CBLA and firearms/toolmarks as prac-
ticed. As they existed in both domains for their decades of prac-
tice, neither proficiency testing nor claimed “validation” studies 
had, nor have, external validity, do not measure what they 
purport to measure, and are virtually irrelevant to any particu-
lar judicial proceeding. Courts have routinely conflated profi-
ciency testing with validation studies. The testing and studies 
have been represented by practitioners and perceived by jurists 
as indicia of both the validity of underlying practice theory and 
of rates of practice error because practitioners and jurists are 
unable to deconstruct, evaluate, and discern the flawed designs 
of their experiments. The tests and studies are so flawed in 
design and interpretation that, among numerous other flaws, 
they do not effectively capture error rates in real-world case 
work, yet they are presented to courts as support for exagger-
ated claims of inferential certainty and near-zero rates of prac-
tice error (Spiegelman and Tobin 2013). CBLA was eventu-
ally discontinued not because of recognition of the misguided 
claims from the domain’s purported validation studies or of the 
flawed design of experiments, which never tested the under-
lying hypotheses (premises), but rather because of the non-
scientific, much-easier-to-understand, absence of research into 
retail product distribution. Critics eventually researching the 
phenomenon decades after CBLA had been admitted in judi-
cal proceedings discovered astonishing saturations of same-
composition product in localized regions of the U.S., dealing a 
fatal blow to its forensic utility (probative value) for solving local 
crimes (Cole et al. 2005).

In the domain of firearms/toolmarks identification, practi-
tioners remain intractable even after years of critical scholarly 
papers, ad hoc committees of the National Academy of Sciences 
(NAS), position statements from the U.S. Department of 
Justice (USDOJ), a proxy for the FBI and ATF laboratories given 
that USDOJ does not have its own crime laboratory), National 
Commission on Forensic Science (NCFS, Cooley 2016), which 
operates under the aegis of the National Institute of Standards 
and Technology (NIST), the President’s Council of Advisors on
Science and Technology (PCAST), affidavits from scientists and scholars, and courtroom challenges by mainstream scientific experts. These entities have articulated serious deficiencies and flaws in purported validation studies, claimed error rates, expressions of certainty, and proficiency testing in the field of firearms identification, resulting in some restrictive judicial rulings, yet none have had any discernible effect on the practice other than to further entrench it. In court testimonies, expert witnesses steadfastly continue to opine individualizations (specific source attributions) with exaggerated and unsupported levels of confidence. Cultural intransigence and lack of scientific comprehension have conjointly conspired to prevent adaptation to exhortations from the most respected voices of both the true scientific community and judicial authority, as evidenced by current courtroom testimonies. Even today, practitioners reject admonishments from the NAS, NCFS (NIST), PCAST, and USDOJ, claiming that none of the bodies have the “training, skill, and experience” to recognize uniqueness when they see it, notwithstanding that the premise of uniqueness has never been established to exist (National Research Council 2009). Although “training and experience” can be incalculably useful for expert testimony in general, it should not be the sole basis for opinions underlying forensic practice, as it is for the forensic practice of firearms/toolmarks identification. As noted by Thornton and Peterson:

Experience is neither a liability nor an enemy of the truth; it is a valuable commodity, but it should not be used as a mask to deflect legitimate scientific scrutiny, the sort of scrutiny that customarily is leveled at scientific evidence of all sorts. To do so is professionally bankrupt and devoid of scientific legitimacy, and courts would do well to disallow testimony of this sort. Experience ought to be used to enable the expert to remember the when and the how, why, who, and what. Experience should not make the expert less responsible, but rather more responsible for justifying an opinion with defensible scientific facts. (Thornton and Peterson 2002, pp. 17–18)

The President’s Council of Advisors on Science and Technology (PCAST) has also confirmed what should be axiomatic: that “training and experience” are inadequate foundations for scientific validity and reliability of specific source attributions in firearms identification forensic practice (PCAST 2016).

On a more global level, a major consequence of the lack of true (mainstream) scientific culture of practitioners and jurists is a critical absence of understanding of proper designs of experiments, of the inherently probabilistic nature of the forensic practices, and of a crucial need for cross-discipline scientific input into development/rehabilitation of the practices. For the many decades of judicial application, the absence of true (mainstream) scientific oversight was largely attributable to a lack of extrajudicial interest in each practice. That fact effectively insulated the working hypotheses of the practices from falsifiability, a critical cornerstone of the scientific method. The insular nature of each of the communities additionally circumvented another critical component of proper scientific validation: reproducibility, the characteristic of replication by external experimenters. For CBLA, even had mainstream scientists had any interest in comparing bullet lead compositions, few, if any, had access to a nuclear reactor, and it is unrealistic to expect that, for firearms identification research, mainstream scientists could obtain proper samples of various firearms and an expensive comparison microscope to perform any comprehensive or meaningful studies or testing in firearms identification pattern-matching practice.

Finally, the user-community (the judiciary) has historically been ill-equipped by academic background to serve as gatekeepers in matters of scientific validity and reliability. Although it has been asserted in numerous judicial opinions that the adversarial process of the judicial system serves to flush out and reject bad science, as observed by the National Research Council of the National Academy of Sciences: The adversarial process relating to the admission and exclusion of scientific evidence is not suited to the task of finding “scientific truth.” The judicial system is encumbered by, among other things, judges and lawyers who generally lack the scientific expertise necessary to comprehend and evaluate forensic evidence in an informed manner, trial judges (sitting alone) who must decide evidentiary issues without the benefit of judicial colleagues and often with little time for extensive research and reflection, and the highly deferential nature of the appellate review afforded trial courts’ Daubert rulings. Given these realities, there is a tremendous need for the forensic science community to improve. Judicial review, by itself, will not cure the infirmities of the forensic science community.  

The bottom line is simple: In a number of forensic science disciplines, forensic science professionals have yet to establish either the validity of their approach or the accuracy of their conclusions, and the courts have been utterly ineffective in addressing this problem. For a variety of reasons—including the rules governing the admissibility of forensic evidence, the applicable standards governing appellate review of trial court decisions, the limitations of the adversary process, and the common lack of scientific expertise among judges and lawyers who must try to comprehend and evaluate forensic evidence—the legal system is ill-equipped to correct the problems of the forensic science community.  

It is difficult to dispute that the dominant common denominator of flawed forensic practice proliferation has been, and remains, the absence of adequate scientific ethos of practitioners developing, or jurists adjudicating, scientific validity and reliability of the practices.

Funding

This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Balke, R. (2015, July 22), “A High Ranking Obama Official Just Called for the ‘Eradication’ of Bitemark Evidence,” Washington Post. Available at https://www.washingtonpost.com/news/the-watch/wp/2015/07/22/a-high-ranking-obama-official-just-called-for-the-eradication-of-bitemark-evidence/  

Bush, M. A., Bush, P. J., and Sheets, H. D. (2011), “Statistical Evidence for the Similarity of the Human Dentition,” Journal of Forensic Sciences, 56, 118–123. 

21 Report of the NRC/NAS, “Strengthening Forensic Science in the United States: A Path Forward” (2009), at S-9; Fn. 19, supra.

22 Ibid, at 1-14.
Cole, S. A., Tobin, W. A., Burgess, L., and Stern, H. (2005), “A Retail Sampling Approach to Assess Impact of Geographic Concentrations on Probative Value of Comparative Bullet Lead Analysis,” Law, Probability & Risk, 4, 119–216. [3,9]

Cooley, C. M. (2016), “Initial Draft: Recommendations to the Attorney General Regarding Use of the Term ‘Reasonable Scientific Certainty,’” National Commission on Forensic Science. Available at https://www.justice.gov/nca/file/816021/download. [9]

FBI (2005), “FBI Laboratory Announces Discontinuation of Bullet Lead Examinations,” Press release September 1. Available at https://archives.fbi.gov/archives/news/pressrel/press-releases/fbi-laboratory-announces-discontinuation-of-bullet-lead-examinations [3]

Franco, A., Willems, G., Souza, P. H. C., Geertruida, E., Bekkering, G. E., and Thevissen, P. (2014), “The Uniqueness of the Human Dentition as Forensic Evidence: A Systematic Review on the Technological Methodology,” International Journal of Legal Medicine, 1–7. [4]

Grissom, B. (2015, November 17), “Arguments Over Bitemarks get Testy at Texas Forensic Science Commission Meeting,” The Dallas Morning News. Available at http://trailblazersblog.dallasnews.com/2015/11/arguments-over-bite-marks-get-testy-at-forensic-science-commission-meeting.html [3]

Hsu, S. S. (2016), Available at https://www.washingtonpost.com/local/public-safety/va-exoneration-underscores-moonscape-challenges-to-bite-marks-evidence/2016/04/08/55b4f98e-8f9a-11e5-886f-a037db23301_story.html [3]

Imwinkelreid, E. J., and Tobin, W. A. (2003), “Comparative Bullet Lead Evidence (CBLA): Valid Evidence or Ipse Dixit?” Okla. City Univ. LR, 28, 43–72. [3]

Innocence Project (2016), Innocence Project: The Cases. Available at http://www.innocenceproject.org/cases-false-imprisonment/front-page#t=0&c=4=Exonerated+by+DNA [3]

National Research Council (2009), Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: National Academic Press. [3,10]

President’s Council of Advisors on Science and Technology (2016), “Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods.” September 2016, Available at https://www.whitehouse.gov/administration/eop/ostp/pcast [7,10]

Rawson, R. D., Ronald, K. O., Kinard, G., Johnson, J., and Yantis, A. (1984), “Statistical Evidence for the Individuality of the Human Dentition,” Journal of Forensic Sciences, 29, 245–253. [5]

Spiegelman, C., and Tobin, W. A. (2013), “Analysis of Experiments in Forensic Firearms/Toolmarks Practice Offered as Support for Low Rates of Practice Error & Claims of Inferential Certainty,” Law, Probability & Risk, 12, 115–113. [9]

Thornton, J. I., and Peterson, J. L. (2002), “The General Assumptions and Rationale of Forensic Identification” in Science in the Law: Forensic Science Issues, eds. D. Faigman, D. Kaye, M. Sacks, and J. Sanders. Eagan, Minnesota: West Publishing Group. ISBN: 0-314-26326-8. [7,8]

Tobin, W. A., and Blau, P. J. (2013), “Hypothesis Testing of the Critical Underlying Premise of Discernible Uniqueness in Firearms/Toolmarks Forensic Practice,” Jurimetrics J., 53, 121–146. Available at http://ssrn.com/author=1521077 [8]

Tobin, W. A., and Duerfeldt, W. (2002), “How Probative is Comparative Bullet Lead Analysis?” Criminal Justice, 17, 26–34. [3]

Tobin, W. A., and Monson, K. L. (1989), “Collapsed Spring Observations in Arson Investigations: A Critical Metallurgical Evaluation,” Fire Technology, 25, 317–335. [5,6]

Turvey, B. E., and Colley, C. M. (2014), Miscarriages of Justice, San Diego, CA: Academic Press, pp. 416. [3]