Shooting blanks
The experiments justifying firearm/toolmarks

Cliff Spiegelman, Texas A&M University
And
Bill Tobin, Forensic Engineering International

Firearm/toolmarks

- Firearm toolmark examinations and comparisons are often used in investigations of homicides involving a firearm, spent bullets and/or cartridge cases that are recovered from crime scenes. Most frequently, one or more firearms are recovered during investigation of a shooting incident and typically submitted for forensic comparisons with bullets and/or cartridge cases recovered from the scene. The forensic practice used to associate or eliminate a particular firearm as the murder weapon is based on comparisons of characteristics imparted to bullets and cartridge cases during cycling of a cartridge through the gun, and is known as firearm/toolmarks examination.

Typical testimony

- If it is concluded that the submitted weapon "matches" the crime scene bullets, the firearm/toolmark examiner typically testifies at trial that the crime scene bullets were fired from the gun to the exclusion of all other possible weapons, although sometimes "to a practical certainty."

Six Papers Typically Presented in Court as Justification of Firearm/Toolmark Statements of Certainty

- The six papers are Brundage (1998), Bunch and Murphy (2003), DeFrance and Arsdale (2003), Orench (2005), Smith (2004), and Thompson and Wyant (2003). We only discuss the first two of the six papers were selected as representative for discussion. Two recent National Academy of Science reports have concluded that there is no statistical foundation for such firearm toolmark individualizations. Notwithstanding, prosecutors use these six papers in their most often successful attempts to convince judges that the scientific basis for such firearm toolmark examiner testimonies is well grounded and accepted in the scientific community.

Bad Experiments Can Take a While to Uncover

- A famous example of poor experimentation is the purported phenomenon of 'polywater.' The example is chosen because large portions of the world's scientific community entertained claims made for polywater. A Russian scientist, Nikolai Fedyakin, showed that repetitive filtering of water through quartz capillary tubes produced water with much lower freezing points and much higher boiling points. In the 1960s, there was full-scale scientific effort to confirm the existence and properties of polywater. Papers in prestigious journals such as Science ostensibly confirmed its existence. The theory triggered fear of a doomsday scenario because it was thought that if polywater came in contact with oceans, they would acquire properties of polywater. Many in the general public were aware of the claims for polywater in the 1970s. As it turned out, 'polywater' was caused by sloppy experimentation with contaminants that had leached into normal water. That is, 'polywater' is dirty water. See Epstein (1998), and Franks (1983) for more detail.

More on Polywater

This example is notable for a number of reasons. In addition to the poor experimental design, it is notable that only through many good experiments of proper experimental design by skeptics was widespread acceptance of polywater overcome. Sometimes unsupported scientific claims take on a life of their own and become entrenched in the popular culture or judicial community, requiring many years to dispel, even when widespread segments of the scientific community are involved.

- Unlike the polywater example, few in the scientific community are actively involved in firearm toolmarks. According to the National Academy of Sciences, critical premises underlying the practice have not been scientifically established. On the other hand, firearm toolmark examiners, in court testimony, publications, and trade and professional organizational guidelines, disagree with the NAS findings and claim that specific source attributions (individualizations) can be inferred from comparisons of bullets and/or cartridge cases found at crime scenes with a single firearm to the exclusion of all other weapons (or to a practical certainty), even without having examined any other specimen in the possible sample space.

Most of what follows is suitable for AP Statistics Courses

- The statistical sophistication needed to understand most of the highlighted experimental issues is an AP statistics course or a typical one-semester statistics course for non-majors. The remaining issues can be well understood by any student who has learned mixed model ANOVA. Students may wish to read the classic reference Box, Hunter, and Hunter (2005), or read Kuehl (2000), a book that is used to teach graduate level experimental design at Texas A&M University for additional background. The first reference is suitable for students nearing completion of an AP statistics course. In order to appreciate the scientific underpinnings of the experiments discussed, it is useful to understand what a toolmark is, what a firearm toolmark is, different forms they take, and how, in general terms, matches are claimed by examiners.

How the two papers will be reviewed

- For both papers that we review, we will present the stated goal of the experiment (if any), hypotheses tested, controlled factors, factors not considered or discussed, sample sizes, author conclusions, and a critique of those conclusions. The demonstrated flaws in the two articles reviewed are representative of those existing in the remaining four papers not reviewed for this article due to space limits. The list of factors not considered or discussed by the authors is not necessarily exhaustive.

1. "A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases" by Bunch and Murphy (2003).

- The stated hypotheses are, "1) that marks imparted to cartridge cases from different guns rarely if ever display sufficient agreement to lead qualified firearms examiners to conclude the specimens were fired from the same gun and 2) that marks imparted to cartridge cases from the same gun will rarely if ever lead a qualified firearms examiner to conclude the specimens were fired from different guns." We rephrase these in traditional statistical framework. Due to the ambiguous wording of the authors' proposition "rarely if ever", the traditional restatement of their hypotheses also carries the ambiguity.

Restated Hypotheses

H_{01} : Marks imparted to cartridge cases from different guns not-infrequently or sometimes display sufficient agreement to lead qualified firearms examiners to conclude specimens were fired from same gun.

H_{a1} : Marks imparted to cartridge cases from different guns rarely if ever display sufficient agreement to lead qualified firearms examiners to conclude specimens were fired from same gun.

And,

H_{02} : Marks imparted to cartridge cases from same gun will not-infrequently or sometimes lead a qualified firearms examiner to conclude specimens were fired from different guns.

H_{a2} : Marks imparted to cartridge cases from same gun will rarely if ever lead a qualified firearms examiner to conclude specimens were fired from different guns.

Factors Chosen by Experimenters

- 45 test bullets given to each examiner which were a mix of Glock and other cartridge casings. Examiners were to treat the challenge as casework.
- Factors and sample sizes chosen by experimenters were:
- Ten Glock Pistols (9 MM) with consecutively manufactured breechfaces
- One Baretta model 92F (Luger 9mm)
- One SigSauer model P226 (Luger 9mm)
- Eight FBI firearm/toolmark examiners
- 42 test fires from consecutively manufactured Glocks
- 318 other cartridges were used; it is difficult to know from what weapons the 318 other cartridges were fired

Factors Missing from the ‘Validity Study’

- Factors not considered or discussed, and sample size issues:
- Ammunition type
- Ammunition charge
- Cartridge case hardness
- Primer cup hardness
- Breechface hardness
- Firing pin hardness
- Different batches of Glocks
- Non-Glock firearms“feeds and speeds” of production (or, alternatively, economic conditions of the manufacturing environment)
- Batch and sample not chosen randomly;
- Two 9 mm lugers were not chosen randomly
- Acceptable undersized ammunition
- Brand of ammunition
- Response measures

Factors Missing from the 'Validity Study' (Continued)

- Ids from firing pins
- Ids from ejector marks
- Ids from breechface marks
- Ids from combination of ejector, firing pin, and breechface marks
- Weapon cleanliness
- Participants' experience as toolmark examiners
- Participants asked to handle test as casework, but no measures of effectiveness for this instruction reported
- Break-in period for pistols
- Lubrication regime, present or not
- Condition of lubrication system, nested within lubrication regime
- Fabrication tooling materials, if any
- Fabrication tooling hardness
- Type of workpiece
- Alloy used for workpiece

Factors Missing from the 'Validity Study' (Continued)

- Temper of workpiece
- Microstructure of workpiece
- Finishing processes (assertion only by author that "breechface is unaffected during all remaining operations" but no description as to what constituted "remaining operations" for interpretation of possible metallurgical effects)
- Effect of checking random sample of test cartridges for clear marks by experimenter before proceeding with the experiment was not investigated. This cannot be done in practice, as criminals do not check to insure their cartridges are clearly identifiable before leaving them at crime scenes.

Stated Conclusions

- There were 70 true IDs (positives) that examiners could make, and 100 percent were made. Of 290 true exclusions, examiners made 118, the remainder was declared inconclusive. Thus, authors state specificity percentage as .407. The two propositions tested by this experiment were confirmed, that is, the two null hypotheses were rejected. [No significance level is reported.]

Comments on the experiment and stated conclusions

There was no meaningful SOP or detailed criteria of how matches were made or not. Factors used in the experiment were too few: only three types of weapon with unbalanced numbers of each were used, the type(s) of ammunition used is unknown, and automated toolmark equipment has shown that ammunition brand matters for automated identification, Bachrach (2006), all examiners from one organization, one manufacturing batch used for the ten Glocks, and unknown number of break-in test fires. The fact that examiners were not chosen at random but rather from what is considered to be an elite unit in the FBI Crime lab limits generality. Even if we pretend that the eight examiners are typical, an exact 95% confidence interval for the percentage of examiners that would also be perfect on a similar exam is approximately 63 to 100 percent.

Comments continued

Thus, a fair assessment would be that at least 63 percent of examiners would do as well on this experiment. The typical sample sizes are small, one group of examiners, three types of weapons (two with sample size 1). Thus, the experiment provides little support for the conclusions presented. In addition, considering the fact that the level of most factors was not recorded (*e.g.*, cartridge or breechface hardness, lubrication regime, or chamber pressure), it is difficult to see how different studies, without properly measured factors, can be combined to assess cartridge IDs. In addition, examiners knew they were being tested and, according to well-accepted principles, this creates a challenge to applying results of this study to general toolmark community actual casework. See the first two chapters of Cook and Campbell (1979), and Shadish, Cook, and Campbell (2002) for factors affecting the validity of field experiments. Examiners could estimate how similar markings on test fires were for the different (consecutively manufactured) weapons provided. In essence, they observed what statisticians call 'between-variation', or an intuitive feel for the 'between sums of squares'. This is quite unusual in actual forensic casework.

Paper #2
“The Identification of Consecutively Rifled Gun Barrels” by David Brundage (1998).

- The general issue is to address concern that guns are massed produced and that toolmarks imparted during the manufacture carry over to the bullets fired from them.

Stated purpose

- The stated purpose is: "Whether two or more bullets, fired from consecutively rifled (manufactured) gun barrels, could be associated with the barrel from which they were fired." There are three research questions stated for the paper. 1) Can trained forensic firearm examiners distinguish between two or more multiple gun barrels that have been consecutively rifled? 2) Can firearm examiners in the Illinois State Police Forensic Science Command differentiate individual characteristics of bullets fired from gun barrels that have been rifled in a consecutive manner? 3) Can firearm examiners from nationally accredited forensic laboratories accurately identify bullets that were fired from gun barrels that have been consecutively rifled?

Restated Hypotheses

The three sets of null and alternative hypotheses are:

H_{01} : Trained forensic firearm examiners make errors at least alpha times 100 percent of the time when trying to distinguish between two or more multiple gun barrels that have been consecutively rifled.

H_{a1} : Trained forensic firearm examiners make less than alpha times 100 percent of the time when trying to distinguish between two or more multiple gun barrels that have been consecutively rifled.

H_{02} : Firearm examiners in the Illinois State Police Forensic Science Command make errors at least alpha times 100 percent of the time they attempt to differentiate individual characteristics of bullets fired from gun barrels that have been rifled in a consecutive manner.

H_{a2} : Firearm examiners in the Illinois State Police Forensic Science Command make errors less than alpha times 100 percent of the time they attempt to differentiate individual characteristics of bullets fired from gun barrels that have been rifled in a consecutive manner.

Restated Hypotheses (Con't)

H_{03} : Firearm examiners from nationally accredited forensic laboratories make errors at least alpha times 100 percent of the time when they identify bullets that were fired from gun barrels that have been consecutively rifled.

H_{a3} : Firearm examiners from nationally accredited forensic laboratories make errors less than alpha times 100 percent of the time when they identify bullets that were fired from gun barrels that have been consecutively rifled.

Factors considered or discussed, and sample sizes

One nine-millimeter Ruger p-85 semiautomatic pistol,

Ten consecutively manufactured barrels,

Barrels "rough cast",

Two lots of ammunition used: Winchester/Olin supplied one lot, manufacturer of second lot not stated,

1200 test bullets fired,

Each bullet was fired from one of ten consecutively manufactured barrels,

Each test set consisted of 15 randomly chosen bullets stratified to insure there was at least 1 bullet from each of ten barrels,

Test packets each containing 35 bullets,

20 test standards (2 from each barrel, not stated but reasonably inferred) and 15 questioned bullets.

Some microscope types

Numbers of participants by city, state

Counts for sworn officers and non-sworn officers

Three types of lighting used are characterized

28/30 examiners in study routinely examine non-firearms evidence

Summary statistics on various professional organizations to which participants belong

Average time for test completion (nine hours)

Factors considered or discussed, and sample sizes (Con't)

- Thirty participants,
- Participants chosen from laboratories accredited by American Society of Crime Laboratory Directors-Laboratory accreditation board (ASCLD-LAB)
- Test packets mailed to participants: pretest administered to five accredited laboratories that did not participate in the final test. In the pretest, a labeling error was found where two sets of bullets were provided for one barrel and none were provided for another. It is clear that these errors were corrected for the final test.

Factors not considered or discussed in described experiment

- Barrel hardness
- Bullet hardness,
- Presence of coatings or jackets,
- Appropriate undersized ammunition,
- Manufacturer of all ammunition,
- Non-rough cast barrels,
- Regime of lubrication, present or not
- Condition of lubrication, nested in lubrication regime
- Barrel alloy,
- Ammunition charge,
- Bullet speed,
- Temperature and expansion coefficients for barrels and ammunition ,

Factors not considered or discussed in described experiment (Con't)

- Lubrication regime operative, if any,
- Non-pristine condition bullets (condition not stated but assumed from typical standard testing protocols),
- Other firearm types,
- Participants not randomly selected,
- And mechanism for selection from certified labs not discussed,
- No controls over group collaboration,
- Examiner level of experience not provided,
- And each weapon was broken-in test fired using fully loaded magazine at factory. (This atypical break-in may well have provided some or all of the marks used to identify barrels.)
- Barrels not randomly selected and manufacturer knew chosen weapons would be used in experiment,
- Unknown if participants treated experiment as casework.

Stated conclusions

- “The project verifies the three research questions regarding the identification of consecutively rifled gun barrels. The results also demonstrate that, on a national level, properly trained examiners can distinguish two or more bullets fired from such barrels. Furthermore, they can accurately differentiate the individual characteristics of test shots from consecutively rifled barrels. This [sic] data also shows that not only are consecutively rifled gun barrels different from each other, but [sic] are unique and can be differentiated.”

Comments on the experiment and stated conclusions

- There was no meaningful SOP or detailed criteria of how matches were made or not. The factors used in the experiment were too few: one type of weapon, possibly two types of ammunition, one manufacturing period, method of manufacture, and rifling for the barrels.) If we pretend that the thirty participants were randomly chosen, then using an exact 95% binomial confidence interval, we would expect between 88.5 percent and 100 percent of equally qualified examiners to do as well for this particular set of test parameters. It is unknown what percentages of the participants are from Illinois.

More comments on conclusions

- Thus, making a separate statement about Illinois participants cannot be supported by the published information. With the typical sample size of one for the many factors, such as type of weapon used in this experiment, exact confidence intervals for the percentage of weapons for which the results hold range from 2.5 to 100 percent with 95 percent confidence. Thus, the experiment provides little support for the hypotheses tested and conclusions presented. The claim of uniqueness of the consecutively manufactured barrels cannot reasonably be made based upon a comparison of test fires from ten barrels.