Reproducibility of Bullet Comparison Conclusions

By: Brandon A. Best and Elizabeth A. Gardner, University of Alabama at Birmingham, Birmingham AL

ABSTRACT

Ten test kits, comprised of damaged and undamaged bullets fired from ten consecutively manufactured barrels, were evaluated by a total of 20 participants with each test kit being examined by two participants for a total of 300 comparisons. Two examiners of varying experience levels analyzing the same test kit reported the same conclusion in 77% (232/300) of the comparisons. Of the 68 differing conclusions between the two examiners for the same test kit, 76% (52/68) reported identification versus inconclusive, 21% (14/68) reported elimination versus inconclusive, and 3% (2/68) reported identification versus false elimination on individual characteristics. This shows that results were reproducible 77% of the time between two examiners.

Introduction

Bullet comparisons are typically conducted using a method known as pattern matching. In this method two fired bullets are placed on a comparison microscope, one on each stage, and a trained examiner compares the striations engraved on the bullet by the barrel, looking for a pattern of agreement between the two sets of striations. This method has been used for over 100 years and has been established as a valid method many times over [1-6]. However, since pattern matching is a cognitive skill, it may be affected by an examiner's training and experience. Examiners with different levels of experience and training could theoretically analyze the same evidence and come to different conclusions. This work is focused on establishing to what extent two examiners will concur or differ in conclusions about the same bullet evidence and if different levels of experience are a factor in those differing conclusions. This article expounds on data related to reproducibility of conclusions from a previously conducted validation study [1].

Methods and Materials

Fifty test kits were created and distributed to examiners as part of a study of ten consecutively manufactured barrels [1]. Each test kit contained 15 comparison sets comprised of ten true identifications and five true eliminations. Ten of these test kits each went out to two different examiners for comparison. Twenty participants, or ten pairs, were asked to compare known bullets to the questioned bullets and

render a conclusion using a variation of the AFTE Range of Conclusions, in which the three types of inconclusive responses were pooled into a simplified single inconclusive category [7]. The participants were selected on a voluntary basis from the AFTE membership. Five of the bullets, selected at random, from each test kit were damaged by firing them through or into common materials a bullet might encounter at a crime scene. These materials were drywall, glass, metal, wood, and Kevlar.

The inclusion of damaged bullets was to replicate real casework as close as possible. One comparison from each kit included a questioned bullet that was fired from a firearm with all the same class characteristics as the consecutive manufactured barrels except for differences in land and groove impression measurements. To further replicate real casework, participants were asked to use their laboratories' policies and procedures, to include any quality control measures. The results of each pair of participants was compared to determine how many conclusions would be the same and how many would be different for the same bullet evidence.

Results and Discussion

Consistency of Conclusions

Ten test kits were examined by 20 participants for a total of 300 comparisons, or 150 pairs of comparisons. In each case, the same test kit was completed by two different participants with different experience levels and from different agencies. The two participants came to the same conclusion 77% (232/300) of the time. Of the 68 comparison results that differed, 76% (52/68) were with one examiner reporting

inconclusive and the other examiner reporting identification for the same comparison set **(Table 1)**. On the other side of the spectrum, 21% (14/68) of the comparisons had one examiner report elimination while the other examiner reported an inconclusive result. Only 3% (2/68) of the comparisons had a stark difference in results with one examiner reporting an identification while the other examiner reporting a false elimination based on individual characteristics. The mean for reproducible results was 11.6 comparisons out of the 15 total comparisons per test kit, with 14 reproducible results being the high and 9 being the low. The average years of experience was ten years and the average difference in years of experience between participants with the same test kit was 11 years.

observed: one from a damaged bullet comparison and one from an undamaged bullet comparison.

On average, examiners with less experience did not report inconclusive results more than experienced examiners. The number of conclusive vs. inconclusive results reported may also be a factor of laboratory policy. Some laboratories do not allow their examiners to make identification conclusions if the suspected firearm is not available for examination, which leads to a potential elevation in inconclusive results. **(Figure 1-Kits G and I)**. Kits G and I were excluded as outliers for determination of experience level due to the amount of inconclusive results reported.

Damaged Bullet Comparison Same Result	86% (84/98)
Damaged Bullet Comparison Different Result	14% (14/98)*
Undamaged Bullet Comparison Same Result	73% (148/202)
Undamaged Bullet Comparison Different Result	27% (54/202) *
Total Same Result	77% (232/300)
Total Inconclusive vs. Identification	76% (52/68)
Total Inconclusive vs. Elimination on Individual Characteristics	9% (6/68)*
Total Inconclusive vs. Elimination on Class Characteristics	12% (8/68)
Total Identification vs. Elimination on Individual Characteristics	3% (2/68)*
*1 False Elimination.	

Table 1: Summary of Results

Experience Level

The number of conclusive results, identification or elimination, reported was more dependent on the quality of the bullet comparisons and level of damage to the bullets than years of experience by the participants. The number of consistent results between the two examiners per test kit was higher for damaged bullet comparisons than undamaged bullets with 86% (84/98) and 73% (148/202), respectively **(Figure 1)**. However, this was due to the higher rate of inconclusive results for damaged bullets. Of the 84 consistent result damaged bullet comparisons, 93% (78/84) were inconclusive. In contrast, undamaged bullet comparisons were inconclusive in 49% (72/148) of consistent results between examiners.

Among the consistent conclusive results for pairs of damaged bullet comparisons, 67% (4/6) were correctly identified and two were properly eliminated based on class characteristics. Of the 76 consistent conclusive results for undamaged bullet comparisons, 84% (64/76) were identification results with the remaining 12 being eliminations on class characteristics. Two false eliminations on individual characteristics were

The number of results that were the same between the two participants per test kit was very similar regardless of the difference in years of experience between the examiners **(Figure 2)**. As the difference in years of experience between examiners increases there is not a linear regression of the number of results that are consistent between the two examiners for each test kit. The lowest number of comparison results that were the same between participants was 9 out of 15 with a difference in examiner experience of 18 years and the highest was 14 out of 15 comparisons with 16.5 years difference in experience.

Errors

There were two false elimination errors reported. In the first error, one participant was inconclusive while the other incorrectly eliminated on individual characteristics. In this error there was a difference of 18 years of experience between the two examiners with the less experienced examiner making the false elimination. In the second error, one participant rendered an identification while the other participant incorrectly eliminated on individual characteristics. In this
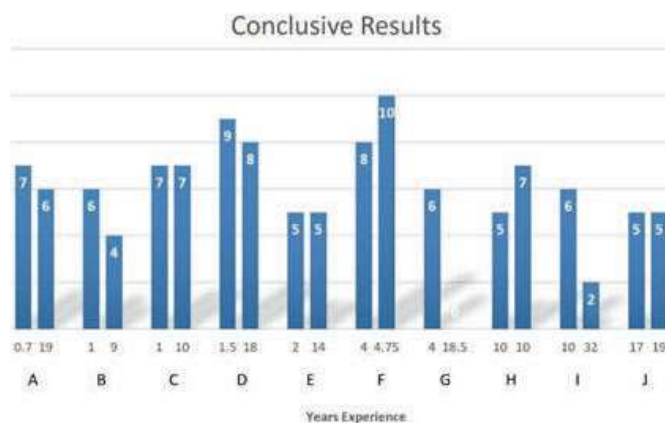
Figure 1: Conclusive results per years of experience for test kits A-J (test kit identifiers changed to protect participants).
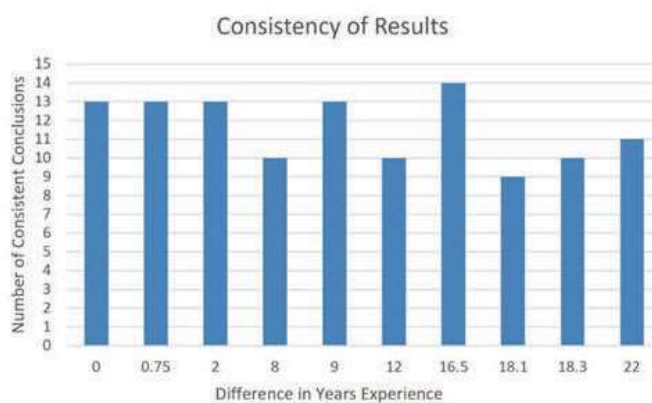

Figure 2: Consistent results per difference in years of experience.

error, there was a difference of 12 years of experience with the more experienced examiner making the false elimination. The difference in these two errors demonstrates that years of experience is not necessarily a factor that contributes to errors. No false identifications were reported for these ten test kits.

Conclusions

It was determined that 77% (232/300) of bullet comparison conclusions were the same between the two examiners. Of the 68 differing comparisons, 76% (52/68) were with one examiner reporting identification while the other reported an inconclusive result. Only 3% (2/68) of the comparisons showed a stark difference in results, with one examiner reporting an identification while the other examiner reporting a false elimination based on individual characteristics.

The number of conclusive results, identification or elimination, were not dependent on the years of experience of the examiner but rather the quality of the striations for comparisons. Bullets with more damage received more inconclusive results regardless of the participant's years of experience. Examiners with more years of experience did not report conclusive results more than those with fewer years of experience and vice versa, on average. Similarly, the difference in years of experience between examiners did not have an impact on the number of comparisons that received the same result for each kit. Two false eliminations on individual characteristics were observed. However, due to the low number of error data points, little weight can be given to how experience level effected error rates in this study. Some work has been conducted on this topic; however, no relationship has been found between years of experience and error rates [8].

Reproducibility of conclusions for bullet comparisons was relatively consistent between participants with the same test kit. Differences in results primarily consisted of one participant reporting an identification conclusion while the second participant reported an inconclusive conclusion. What constituted "sufficient agreement" for one participant may not have been the same level of agreement for another. This could have been due to the condition of the bullets, but also may be due to differences in training, experience, or laboratory policy.

Acknowledgements

References

[1] Best BA, Gardner EA. An Assessment of the Foundational Validity of Firearms Identification Using Ten Consecutively Button-Rifled Barrels. AFTE Journal. 2022;54(1):28-37.
[2] DeFrance CS, Van Arsdale MD. Validation Study of Electrochemical Rifling. AFTE Journal. 2003;35(1):35-37.
[3] Fadul Jr. TG, Hernandez GA, Wilson E, Soitloff S, Gulati S. 2013. An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing Consecutively Manufactured Glock EBIS Barrels with the Same EBIS Pattern. DOJ Document No. 244232.
[4] Hamby JE, Brundage D, Thorpe J. The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries. AFTE Journal. 2009;41(2):99-110.
[5] Monson KL, Smith ED, Peters EM. Accuracy of Comparison Decisions by Forensic Firearms Examiners. Journal of Forensic Sciences. 2022;00: 1-15. https://doi.

org/10.1111/1556-4029.15152

[6] Smith TP, Smith AG, Snipes JB. A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework. Journal of Forensic Sciences. 2016;61(4):939-946.

[7] AFTE Range of Conclusions. AFTE Journal. 1992;24(3).

[8] Bajic SJ, Chumbley LS, Morris M, Zamzow D. Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons. Ames Laboratory-USDOE Technical Report # ISTR-5220. 2020.