9.6.6 NBIDE: Combining N-3D Firing Pin and Breech Face Analyses

The NBIDE Top Ten results in Table 9-2 show that when using the N-3D breech face impression data, 101 of the 108 NBIDE casings had the maximum eight correct matches and the other seven casings each had seven correct matches, included in their corresponding lists of ten most highly correlated compared casings. When combining measures by including all casings that make the Top Ten correlation list of *either* breech face or firing pin, 107 of the 108 NBIDE casings had the maximum eight correct matches. Only one casing (RR #99, Winchester, fired from Sig Sauer 33) had only seven correct matches using either region. The compared casing that was omitted from both these Top Ten lists was RR #34 Remington. Conversely, when Casing RR #34 was the reference casing, its Top Ten list for breech face impressions did include RR #99, albeit in the 10^{th} and final position on the list. One summary statistic of the benefit of using both regions is that the average number of correct matches per Top Ten list improved to 7.99 (out of a maximum 8 correct matches) from 7.94 for breech face alone and 5.63 for firing pin alone for N-3D.

9.7 I-2D Correlation Scores

9.7.1 I-2D Scores of NBIDE Firing Pin Impressions

Figure 9-26 contains the color score matrix of the I-2D correlation scores of the 108 NBIDE firing pin impressions. The I-2D correlations were performed using BrassCatcher Software Version 3.4.5.The I-2D results in this section are based on searches involving the extended NBIDE set of 144 casings, but any results involving the 36 Speer casings were omitted before the analysis. Are any gun and ammunition brand patterns evident?

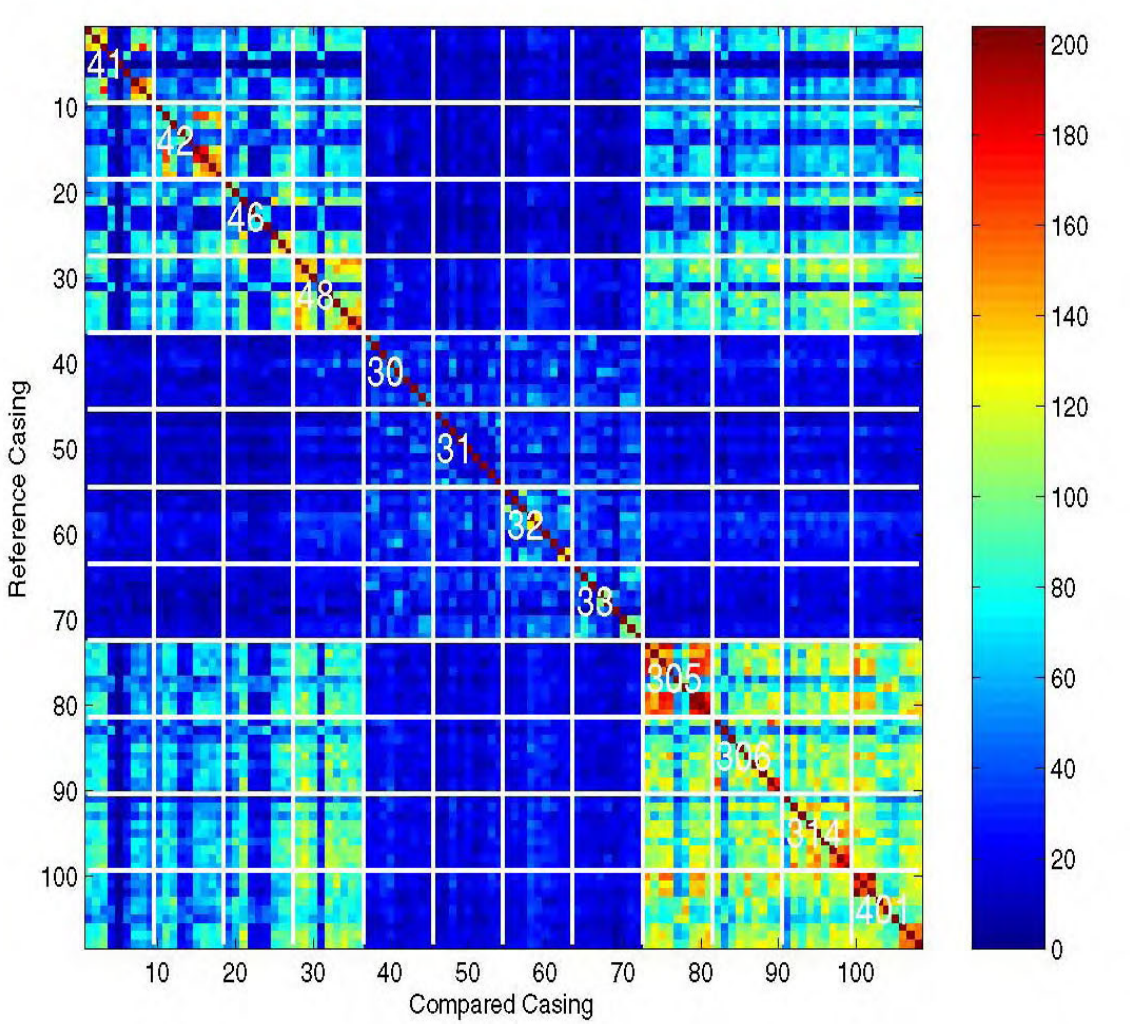

Figure 9-26. Color score matrix of the NBIDE Firing Pin I-2D scores. The color of the pixel in row *I* and column *J* indicates the value of the correlation score between casing *I* as the reference casing and casing *J* as the compared casing. The pixels are ordered by gun ID (Ruger 41, Ruger 42, Ruger 46, Ruger 48, Sig Sauer 30, Sig Sauer 31, Sig Sauer 32, Sig Sauer 33, S&W 305, S&W 306, S&W 314, S&W 401, with Rugers at upper left and S&W's at lower right), within gun by ammunition (1-Win, 2-Rem, 3-PMC), and within ammunition by RR#. The self-correlation scores on the main diagonal were arbitrarily assigned the color of the maximum score present (204).

The Sig Sauers are responsible for the dark blue cross in the middle of the above figure. These guns do not correlate with guns of the other two brands. However, matching scores of the Sig Sauers are also lower than those from the other two gun brands. The opposite is true of the S&W's, in that they have higher matching scores, but also higher non-matching scores, especially with the other S&W's and with Ruger 48. Also, the matching scores for most of the guns appear to be higher if the ammunition brands as well as the guns are the same.

Figure 9-27 depicts histograms of the matching and non-matching scores for the firing pin impressions.

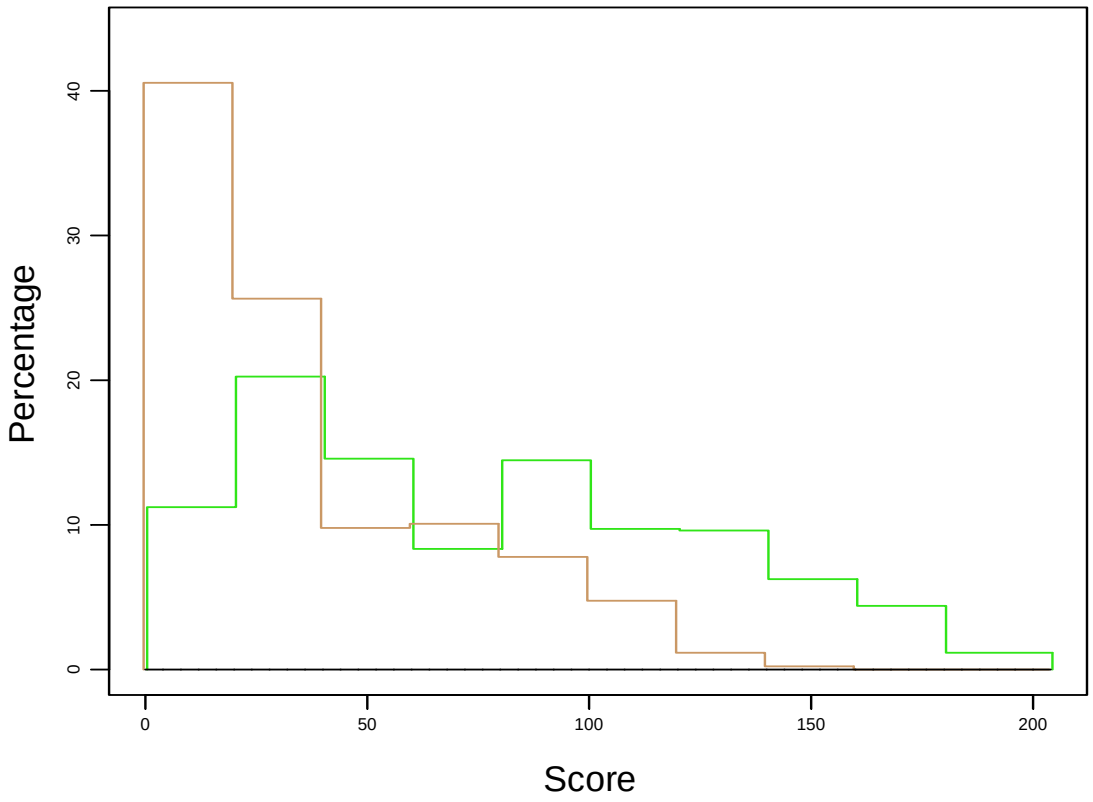

Figure 9-27. NBIDE Firing Pin I-2D data: The green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores.

There is a considerable degree of overlap between the two distributions of scores. Given the gun differences seen in Fig. 9-26, it makes more sense to examine Fig. 9-28 below, which groups these results by reference gun, with the overlap metric *p* given for each grouping. The figure shows some differences between gun brands. Clearly, the Sig Sauers have both matching and non-matching scores of low magnitude, while the S&Ws have matching and non-matching scores of higher magnitude, with the Rugers in between. All guns appear to have considerable overlap.

Table 9-9 contains the statistics for the firing pin impression data ordered by performance of the gun as estimated by the overlap metric. How do the gun brands differ?

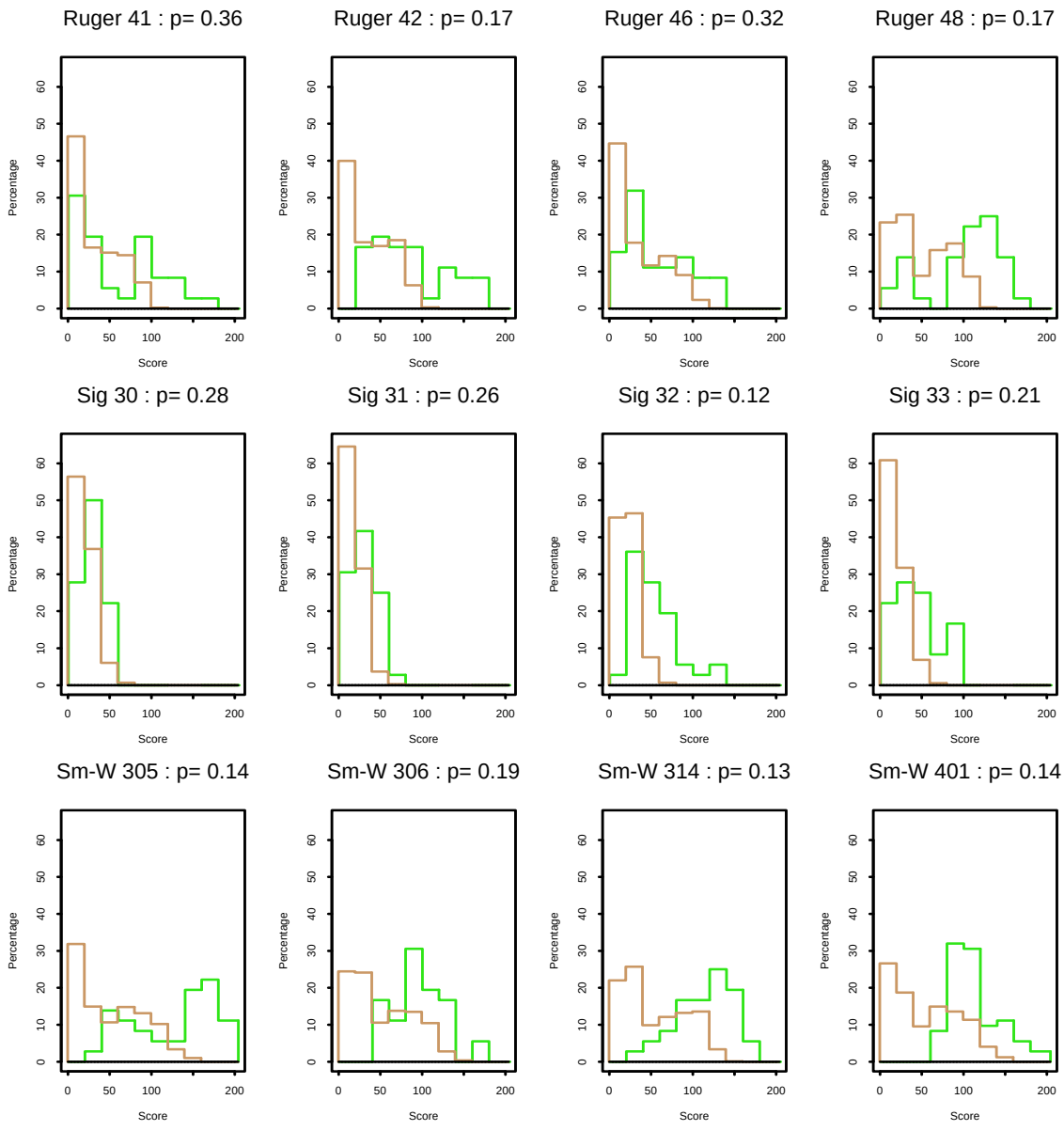

Figure 9-28. NBIDE Firing Pin I-2D data: Above each of the individual plots is a heading with the ID of the reference gun and the overlap metric *p* estimated for that group. In each plot, the green lines depict histograms of the matching scores, while the brown lines depict histograms of the non-matching scores.

Table 9-9. Estimated overlap metric *p* for the I-2D scores of the firing pin impressions of 12 NBIDE guns.

Gun ID	*p*
Sig Sauer 32	0.12
S&W 314	0.13
S&W 401	0.14
S&W 305	0.14
Ruger 48	0.17
Ruger 42	0.17
S&W 306	0.19
Sig Sauer 33	0.21
Sig Sauer 31	0.26
Sig Sauer 30	0.28
Ruger 46	0.32
Ruger 41	0.36
Mean	0.21

There are differences within guns of the same brand. However, the S&W's are on the average doing better than the Sig Sauers. A couple of Rugers are hurt by having some very low matching scores, as can be seen from Figures 9-26 and 9-28.

Given the differences even within the same gun, let's look at Fig. 9-29, which breaks down the data even more finely, by dividing the data into groups by reference casing. Each of the smaller plots depicts for each casing, its I-2D correlation scores with the eight casings fired from the same gun (matches, lower triangles), and the correlation scores with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun number and ammunition of the reference casing.

Most casings appear to have substantial overlaps between matching and non-matching scores. Ruger 48 produces better separation than most except for one casing, which does not correlate well with the other casings fired from that gun. Casing RR 94 fired from Ruger 41 resembles the Sig Sauers in having uniformly low matching and non-matching scores. For each of the 108 casings, an overlap metric for the matching and non-matching correlation scores produced by that casing can be estimated by looking at all pair-wise comparisons between matching and non-matching correlations for each casing. The histogram of Fig. 9-30 shows the empirical distribution of those overlap metrics.

The mean estimated *p* is 0.19, and the median estimated *p* is 0.16. The minimum estimated *p* is 0.04, and around 75 % of the estimates are larger than 0.11. All have a degree of overlap that would produce mistakes in a large database scenario.

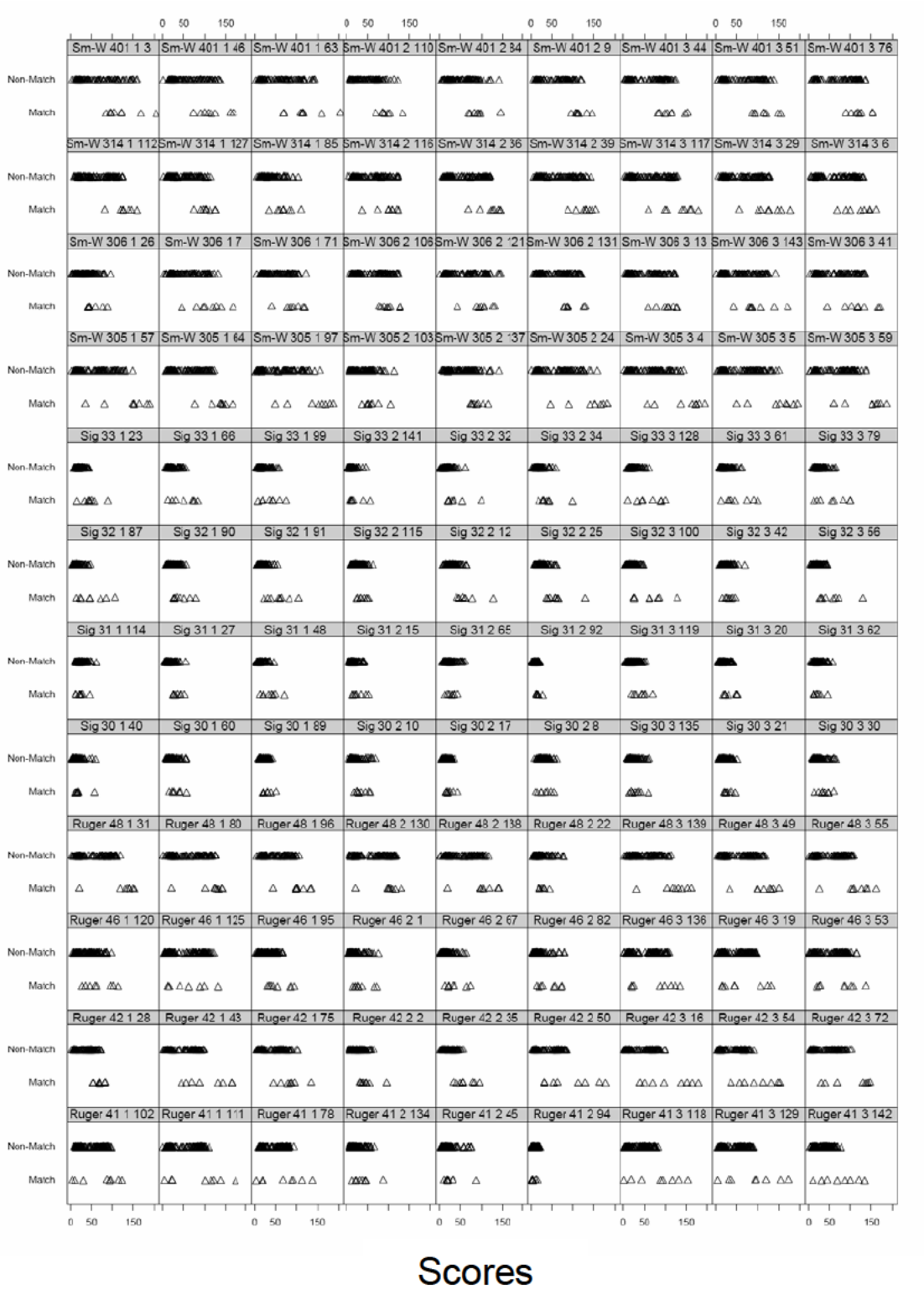

Figure 9-29. NBIDE Firing Pin I-2D data: Correlations, for individual casings, with the eight casings fired from the same gun (matches, lower triangles), and with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID, ammunition (1-Win, 2-Rem, 3-PMC), and RR# of the reference casing.

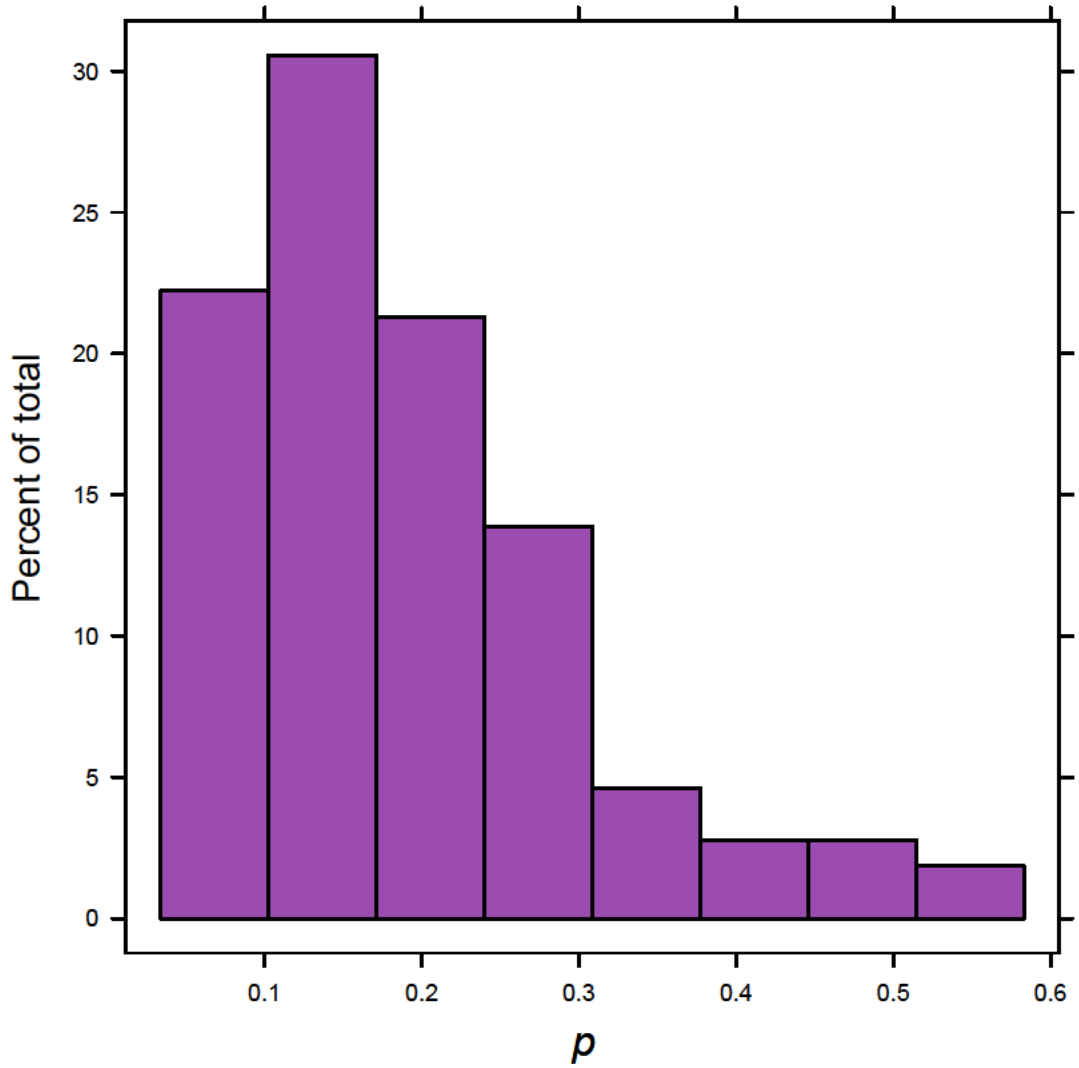

Figure 9-30. NBIDE Firing Pin I-2D data: a histogram of the overlap metric p estimated for each of the 108 NBIDE casings.

9.7.2 I-2D Scores of NBIDE Breech Face Impressions

Figure 9-31 depicts the I-2D scores of the 108 NBIDE breech face impressions in the form of a color matrix. Are the same gun and ammunition brand patterns evident as in the firing pin impressions?

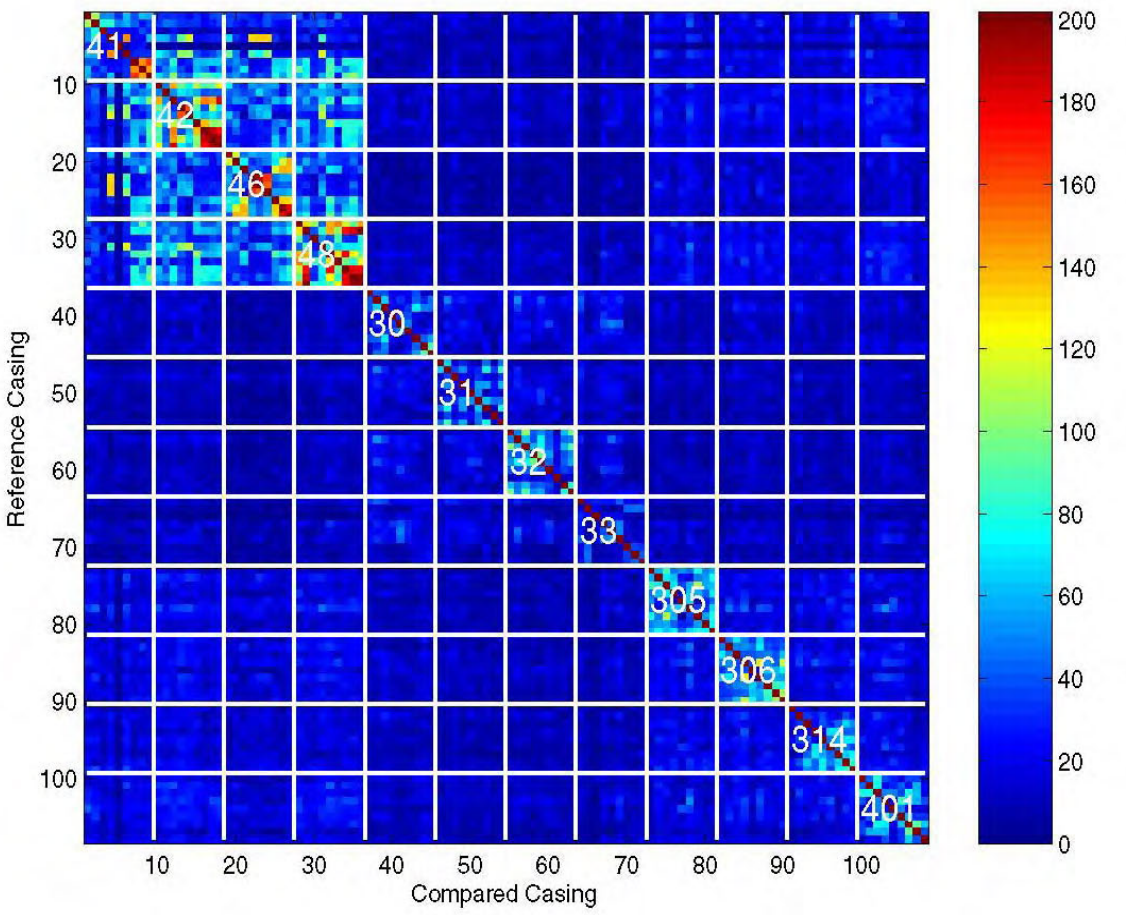

Figure 9-31. Color score matrix of the NBIDE breech face I-2D scores. The color of the pixel in row *I* and column *J* indicates the value of the correlation score between casing *I* as the reference casing and casing *J* as the compared casing. The pixels are ordered by gun ID (Ruger 41, Ruger 42, Ruger 46, Ruger 48, Sig Sauer 30, Sig Sauer 31, Sig Sauer 32, Sig Sauer 33, S&W 305, S&W 306, S&W 314, S&W 401, with Rugers at upper left and S&Ws at lower right), and within gun by ammunition (1-Win, 2-Rem, 3-PMC), and within ammunition by RR#. The self-correlation scores on the main diagonal were arbitrarily assigned the color of the maximum score present (202).

The Sig Sauers again produce lower matching and non-matching scores than do the other gun brands, although the pattern is not nearly as pronounced as for the firing pin impressions. The Rugers have the highest matching scores, especially when both the reference and compared casings are PMCs. However, the highest non-match scores occur when both guns are Rugers; Ruger 41 is an exception to that observation in having several very low matching and non-matching scores.

Figure 9-32 depicts histograms of the matching and non-matching scores for the breech face impressions. What is the degree of overlap or separation between the two distributions?

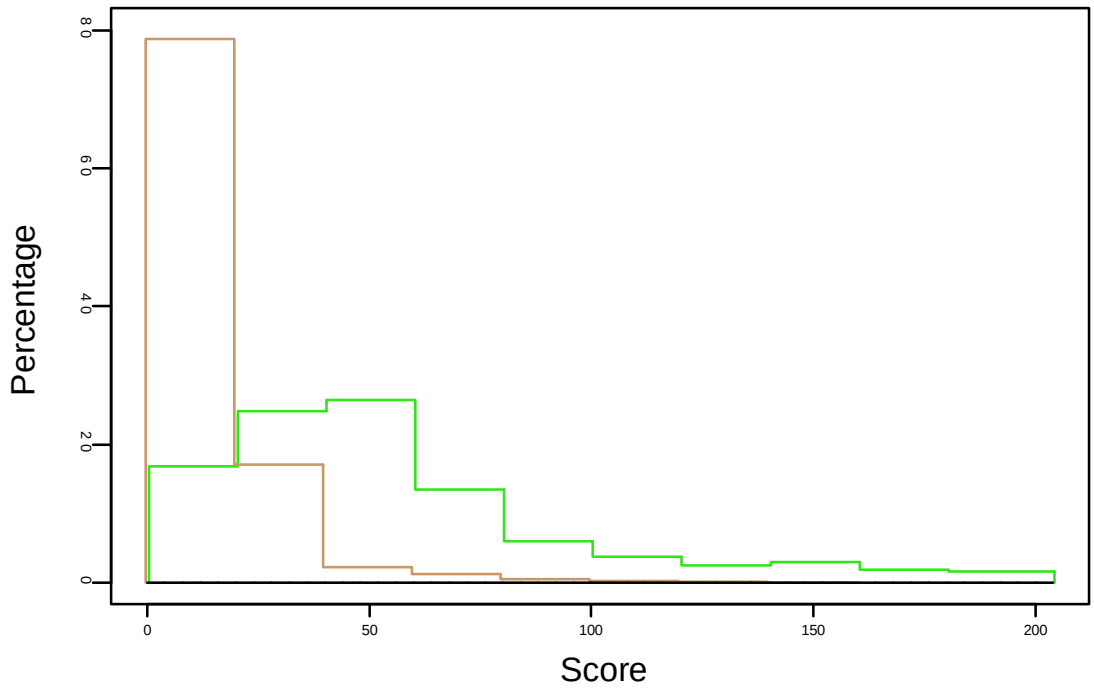

Figure 9-32. NBIDE Breech Face I-2D data: The green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores.

There is a considerable degree of overlap between the matching and non-matching scores, although less than for the firing pin impressions. Given the gun differences seen in Fig. 9-31, it makes more sense to examine Fig. 9-33, which groups these results by reference gun, with the overlap metric *p* given for each grouping. Are the gun brand differences evident?

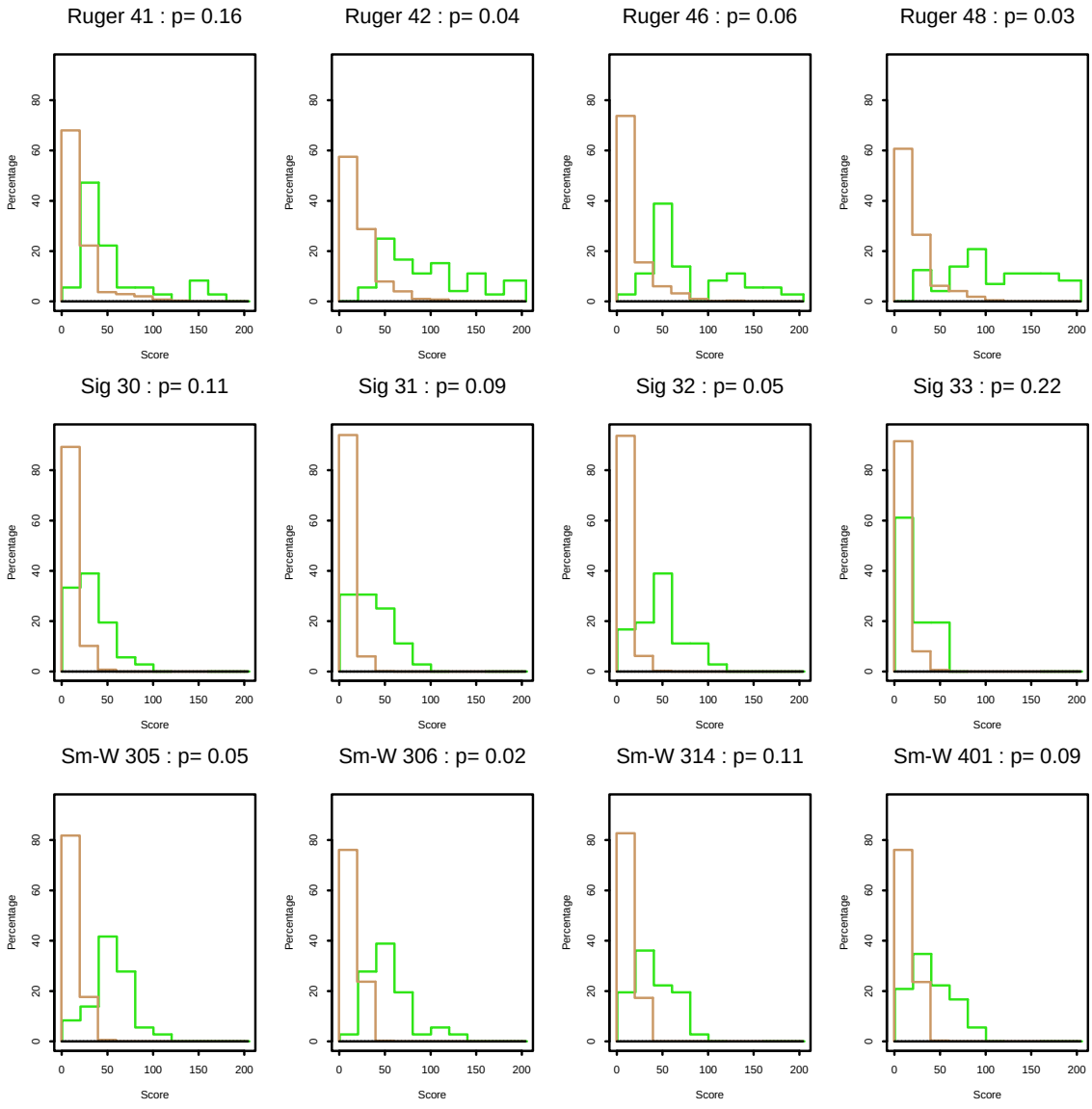

Figure 9-33. NBIDE Breech Face I-2D data: Above each of the individual plots is a heading with the ID of the reference gun and the overlap metric *p* estimated for that group. In each plot, the green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores.

As with the firing pin impressions, the Sig Sauers have both matching and non-matching scores of low magnitude. However, there is less overlap between both matching and non-matching scores than with the firing pin impressions.

Table 9-10 contains the statistics for the breech face data ordered by performance of the gun as estimated by the overlap metric. There is much greater separation between matching and non-matching scores than existed with the firing pin impressions. Also, there are again wide differences within gun brands.

Table 9-10. Estimated overlap metric *p* for the I-2D scores of the breech face impressions of 12 NBIDE guns.

Gun ID	*p*
S&W 306	0.02
Ruger 48	0.03
Ruger 42	0.04
S&W 305	0.05
Sig Sauer 32	0.05
Ruger 46	0.06
Sig Sauer 31	0.09
S&W 401	0.09
S&W 314	0.11
Sig Sauer 30	0.11
Ruger 41	0.16
Sig Sauer 33	0.22
Mean	0.08

Given the differences even within the same gun, let's look at Fig. 9-34, which breaks down the data even more finely, by dividing the data into groups by reference casing. Each of the smaller plots depicts for each casing, its correlations with the eight casings fired from the same gun (matches, lower triangles), and its correlations with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun # and ammunition of the reference casing.

From Fig. 9-34, most but not all casings have some overlap between matching and non-matching scores; however, there is less overlap than is present with the firing pin impression data. For each of the 108 casings, an overlap metric for the matching and non-matching correlation scores produced by that casing can be estimated by looking at all pair-wise comparisons between matching and non-matching correlations for that casing.

Figure 9-35 shows the empirical distribution of those estimated overlap metrics grouped by casing. How does it compare with that of the firing pin estimated scores shown in Fig. 9-30? The estimated overlap metrics are significantly better (closer to 0) than those for firing pin impressions. The mean of the estimated *p* is 0.09, and the median estimated *p* is 0.06. About 75 % of the estimated overlap metrics are larger than 0.027. Seven of the 108 casings have no overlap, leading to estimates of $p = 0$.

Note that because the triangles in Fig. 9-34 have non-zero width, some of the casing plots may give the impression of somewhat more overlap than actually exists. For example, casing RR# 41 fired from S&W 306 (ammunition 3) has no overlap between matching and non-matching scores. Several of those casings with no overlap have little separation between the largest non-matching score and the smallest matching score.

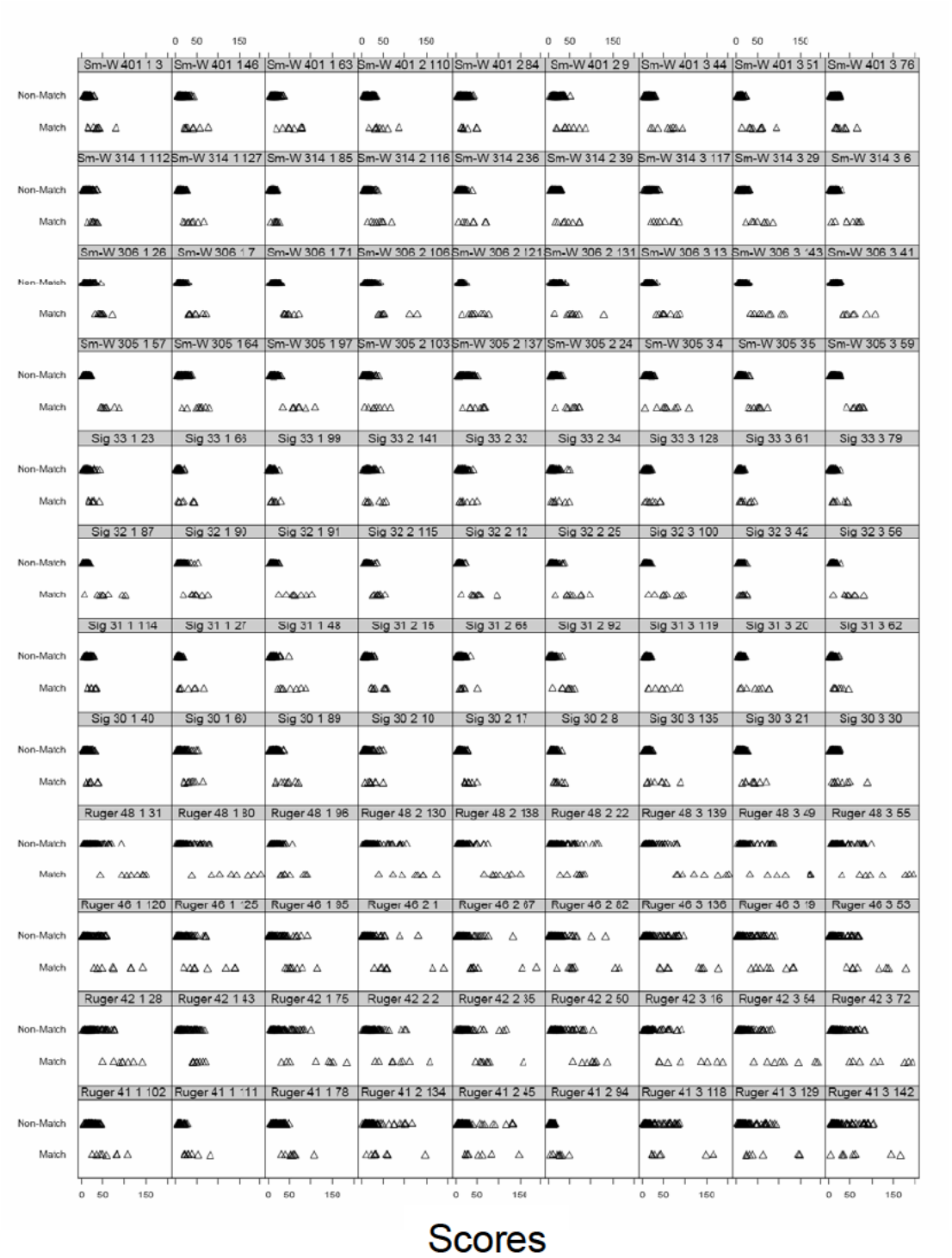

Figure 9-34. NBIDE Breech Face I-2D data: Correlations, for individual casings, with the eight casings fired from the same gun (matches, lower triangles) and with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID, ammunition (1-Win, 2-Rem, 3-PMC), and RR# of the reference casing.

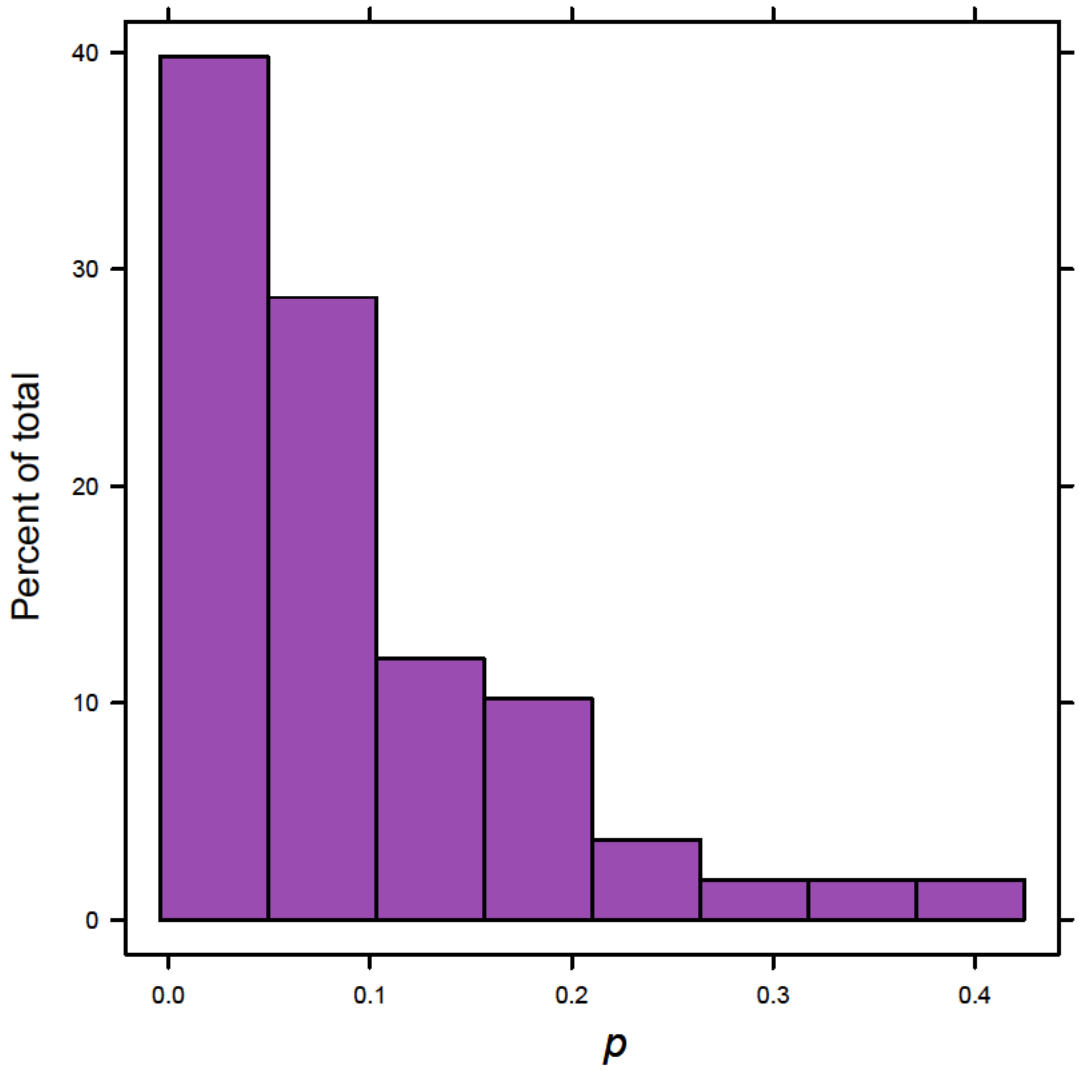

Figure 9-35. NBIDE Breech Face I-2D data: Histogram of the overlap metric *p* estimated for each of the 108 NBIDE casings.

9.7.3 NBIDE: Combining I-2D Firing Pin and Breech Face Analyses

The I-2D Top Ten results in Table 9-2 show that the I-2D breech face impression data produce an average of 5.57 correct matches out of a maximum eight correct matches per Top Ten list. Using the I-2D firing pin impression data yields an average of 3.72 correct matches per Top Ten list. Combining measures by including all casings that make the Top Ten correlation list of *either* breech face or firing pin set yields an average of 6.20 correct matches out of eight, which is certainly better than either region alone, but still does not come close to eliminating the existing coverage mistakes.

9.7.4 Correlation Software and Algorithm Uncertainties

The I-2D correlations of the NBIDE casings were performed in two different ways. First, correlations were performed at the ATF shortly after the I-2D image acquisitions were made. Rankings of the top 30 to 40 breech face and firing pin correlations for each NBIDE casing were calculated and printed. The database for this correlation included all 144 NBIDE test fires and the 70 De Kinder casings. The top results for the NBIDE PMC, Remington, and Winchester ammunition were then entered manually into a spreadsheet and analyzed in a manner similar to that shown in Sec. 9.3 to produce Top Ten results among the 108 NBIDE casings. Afterwards, the data were moved to FTI, Montreal, where the entire collection of 144 NBIDE test fires were analyzed. This calculation produced an array of 144 by 143 I-2D correlation scores, which produced the results in Secs. 9.3 and 9.7.

Although the two correlations produced similar results, there were two types of differences. First, some pairs produced different scores in the ATF and FTI correlations. Although these differences were small in most cases, a few gave significantly different results. For example, the pair of breech face impressions from RR #16 as the reference casing and RR #28 as the compared casing produced an I-2D score of 89 in the ATF correlation but a score of 82 in the FTI correlation. Second, some pairs scoring in the Top Ten in the FTI correlation were completely missing in the ATF correlation.

The first type of difference suggests that there were slightly different versions of software used in the two correlation runs. The second type of difference is ascribed to a coarse filter procedure, which was likely used for the correlation run at the ATF but not for the correlation run at FTI. Because the correlations are generally used with large databases, the I-2D system routinely applies a coarse filter to correlation runs. That is, a preliminary correlation procedure is applied to all the entries in the database. Then for each reference casing only the highest scoring acquisitions, approximately 20 %, are used for a more extensive correlation calculation. The scores from the second correlation are recorded in the printed correlation results, but some of these high scoring pairs may be missing if they do not pass the coarse screening filter test. The frequency of missing entries in the correlation run at ATF was particularly significant when the reference casing was a Sig Sauer test fire because the database there included 36 Sig Sauer test fires from the NBIDE collection and 70 Sig Sauer test fires from the De Kinder collection. Therefore, the probability that a genuine match of casings would fall below the top 20 % (about 44 entries) in the first correlation pass was higher for a Sig Sauer test fire than for a Ruger or S&W, given that there were many similar Sig Sauer test fires in the ATF database.

The software differences discussed above produced differences in the Top Ten scores for the I-2D correlations. Referring to Table 9-2, the ATF average score for breech face impressions was about 5.1 correct matches versus the value of 5.57 given in the Table and derived from the FTI correlations. For the firing pin impressions the average scores were both about 3.7.

A third type of difference in the I-2D scores was also recorded. The matching score for a pair could change depending on which member was used as the reference. However, the differences were insignificant except for a few cases. This asymmetry was also present in the $ACCF_{max}$ scores for the topography correlations.

9.8 NBIDE Topographic Analysis for Gun Brand and Ammunition Effects

The NBIDE-designed experiment with its three gun types and four guns per type allows us to conduct a third kind of analysis that is not available from the De Kinder casings experiment with its single gun type. The experiment enables exploration of ammunition and gun brands on cross correlations. For instance, will casings fired from the same gun be more highly correlated if the ammunition brands are the same? Conversely, will the casings fired from different guns be more likely to be falsely matched if the ammunition are the same brand and/or the guns are of the same model?

For this analysis, we can divide the 108 × 107 casing permutations into six groups shown below with varying degrees of separation. Note that there are 108 × 107 pairs because the $ACCF_{\max}$(A,B) and $ACCF_{\max}$(B,A) are not necessarily the same, so both are included for each combination of casings A and B. This accounts for the final '× 2' in each group enumeration below.

In that regard, we note the following categories of casing pairs:

1) **Same gun, same ammo:**
(12 guns × 3 ammos) × [3 casing combinations × 2]
= 36 × 6 = 216

2) **Same gun, different ammo:**
(12 guns × 3 ammo combinations) × [9 casing permutations × 2]
= 36 × 18 = 648

3) **Different guns of same brand, same ammo:**
(3 ammo brands × 3 gun brands) × [6 gun pairs × 9 casing permutations × 2]
= 9 × 108 = 972

4) **Different guns of same brand, different ammo:**
(3 gun brands × 3 ammo combinations) × [12 gun permutation pairs × 9 casing permutations × 2] = 9 × 216 = 1944

5) **Guns of different brands, same ammo:**
(3 ammo brands × 3 gun brand pairs) × [16 gun permutations] × 9 casing permutations × 2] = 9 × 288= 2592

6) **Guns of different brands, different ammos:**
(6 gun brand permutations × 3 ammo brand pairs) × [16 gun permutations × 9 casing permutations × 2] = 18 × 288 = 5184

The six sub-totals above total to 108 × 107 = 11556.

In this section, we tabulate and explore the averages of the different groups in the first 5 categories above

9.8.1 NBIDE Firing Pin Images: Topographic Analysis of Ammunition Effects

Table 9-11 shows average $ACCF_{max}$ values for various matching casing pairs divided into groups by guns and ammunition combinations. The first four columns of numbers are for the groups where both casings are the same ammunition brand. The last four columns are when the casings are from different ammunition brands.

Table 9-11. Average $ACCF_{max}$ values from firing pin impressions for NBIDE casing pairs fired by the same firearm. All values are expressed in % where 100 % is the value 1.0, indicating perfect correlation.

Gun ID	Same Ammo Brand				Different Ammo Brands			
	Winchester	Remington	PMC	Average	Win-Rem	Win-PMC	Rem.-PMC	Average
Ruger 41	55	27	67	49.7	33	59	33	41.7
Ruger 42	54	66	69	63.0	60	60	66	62.0
Ruger 46	45	24	42	37.0	26	46	25	32.3
Ruger 48	42	27	45	38.0	36	46	35	39.0
Sig Sauer 30	33	24	41	32.7	32	40	37	36.3
Sig Sauer 31	27	27	29	27.7	29	30	30	29.7
Sig Sauer 32	32	53	51	45.3	31	32	51	38.0
Sig Sauer 33	34	42	58	44.7	24	41	33	32.7
S&W 305	48	53	63	54.7	54	54	53	53.7
S&W 306	24	52	37	37.7	27	34	41	34.0
S&W 314	43	33	48	41.3	37	47	37	40.3
S&W 401	45	50	46	47.0	26	38	43	35.7
Average	40.2	39.8	49.7	43.2	34.6	43.9	40.3	39.6

There is obviously a large degree of variability in the category-pair means between gun brands, between guns of the same brand, and even between categories from the same gun. There are ammo effects, but they are not consistent, e.g., lower $ACCF_{max}$ values for Remingtons with Rugers.

The standard deviations of $ACCF_{max}$ values of casing-pair groups, fired by the same gun and having the same ammunition brand, range from 1 % to 23 %, with the very largest standard deviations occurring with the Sig Sauer pairs. The matching casing-pair groups with different ammunition have group standard deviations ranging from 2 % to 20 %, with again the very largest standard deviations occurring with the Sig Sauer pairs.

Table 9-12 shows the average $ACCF_{max}$ values of casings from non-matching guns of the same brand. The columns are organized by ammunition brand combination. The standard deviations of the $ACCF_{max}$ groups in this table range from 3 % to 8 %, with the higher standard deviations occurring with the S&W's.

Table 9-12. Average $ACCF_{max}$ values from firing pin impressions for casing pairs fired by different firearms of the same brand. All values are expressed in %.

	Winchester	Remington	PMC	Win-Rem	Win-PMC	Rem.-PMC
Ruger	22	18	23	20	23	20
Sig Sauer	25	26	24	24	23	25
S&W	24	34	32	26	27	31

Table 9-13 shows the mean $ACCF_{max}$ values from firing pin impressions for casing pairs of the same ammunition brand, but fired from different brands of guns. The standard deviations of the $ACCF_{max}$ groups in this table range from 4 % to 7 %.

Table 9-13. Average $ACCF_{max}$ values from firing pin impressions for casing pairs fired by different brands of guns. All values are expressed in %.

	Winchester	Remington	PMC
Ruger-Sig Sauer	21	21	21
Ruger-S&W	23	22	26
Sig Sauer-S&W	20	26	24

For the large pool of pairs of casings that have both different gun brands and different ammunition brands, the mean of the $ACCF_{max}$ value is 23 % and the standard deviation is 6 %. These appear quite similar to the $ACCF_{max}$ values with the same brand ammunition but different brands of guns.

9.8.2 NBIDE Breech Face Images: Topographic Analysis of Ammunition Effects

Table 9-14 shows average $ACCF_{max}$ values for various matching casing pairs divided into groups by guns and ammunition combinations. The first four columns of numbers are for the groups where both casings are the same ammunition brand. The last four columns are when the casings are from different ammunition brands.

Table 9-14. Average $ACCF_{max}$ values from NBIDE breech face impressions for matching casing pairs fired by the same gun. All values are expressed in %.

Gun ID	Same Ammo Brand				Different Ammo Brands			
	Win.	Rem.	PMC	Ave.	Win-Rem	Win-PMC	Rem.-PMC	Ave.
Ruger 41	78	68	82	76.0	68	76	66	70.0
Ruger 42	76	80	83	79.7	76	78	74	76.0
Ruger 46	51	64	75	63.3	57	64	60	60.3
Ruger 48	81	50	77	69.3	70	80	69	73.0
Sig Sauer30	63	73	64	66.7	58	62	62	60.7
Sig Sauer 31	78	71	69	72.7	74	74	72	73.3
Sig Sauer32	67	74	62	67.7	66	65	67	66.0
Sig Sauer 33	45	55	51	50.3	41	48	40	43.0
S&W 305	51	54	50	51.7	44	52	49	48.3
S&W 306	51	67	70	62.7	54	56	63	57.7
S&W 314	40	57	53	50.0	47	46	50	47.7
S&W 401	62	48	58	56.0	44	56	50	50.0
Average	61.9	63.4	66.2	63.8	58.3	63.1	60.2	60.5

The standard deviations of $ACCF_{max}$ values of those casing pairs from the same gun and the same ammunition brand range from 1 % to 20 %, with the very largest standard deviations occurring with the Sig Sauer pairs. The matching casing pair categories with different ammunition have standard deviations ranging from 5 % to 11 %. The Rugers have higher $ACCF_{max}$ values, and the S&W's have lower $ACCF_{max}$values. Sig Sauer 33 has particularly low $ACCF_{max}$ values.

Table 9-15 shows the average $ACCF_{max}$ values of non-matching casing pairs with common gun brands, organized by ammunition brand combination. The standard deviations of the $ACCF_{max}$ groups in this table range from 3 % to 7 %, with the higher standard deviations occurring with the Sig Sauers and the lower standard deviations occurring with the S&Ws. The one pattern that seems evident is that there are lower $ACCF_{max}$ values between non-matching S&W casings as opposed to non-matching casings of the other two brands.

Table 9-15. Mean $ACCF_{max}$ values from NBIDE breech face impressions for non-matching casing pairs fired by different guns of the same brand. All values are expressed in %.

	Ammunition			Ammunition Pair		
Gun Brand	Winchester	Remington	PMC	Win-Rem	Win-PMC	Rem-PMC
Ruger	22	26	21	24	21	23
Sig Sauer	26	23	24	24	24	23
S&W	19	19	18	19	18	18

Table 9-16 shows the mean $ACCF_{max}$ values of casing pairs of the same ammunition brand, but fired from different brands of gun. The standard deviations of the $ACCF_{max}$ values in this table range from 3 % to 5 %.

Table 9-16. Mean $ACCF_{max}$ values from NBIDE breech face impressions for casing pairs of the same brand fired by different brands of guns. All values are expressed in %.

	Winchester	Remington	PMC
Ruger-Sig Sauer	22	24	20
Ruger-S&W	19	21	19
Sig Sauer-S&W	22	19	20

For the large pool of casing pairs that have both different gun brands and different ammunition brands, the mean of the $ACCF_{max}$ values is 21 % and the standard deviation is 4 %. These values are similar to the $ACCF_{max}$ values with the same brand ammunition but different brands of guns.

Any variabilities due to ammunition are smaller than the variabilities due to guns. We have seen considerable variability even within guns of the same model. In any case, ammunition brand effects are relatively modest compared to gun effects, because the non-matching scores tend to be around the same magnitude regardless of whether the ammunition brands are the same or whether the gun brands are the same. An exception may be the higher firing pin non-matching $ACCF_{max}$ values for Smith&Wesson guns using non-Winchester ammunition. On the other hand, the matching scores vary most according to the gun brand and between individual guns within the same gun brand.

Some analyses of variance found that gun and ammunition effects, as well as interactions, were statistically significant for the matching scores. For the non-matching scores, gun and ammunition effects were also significant. In all cases, the largest effects were the gun effects. Section 10 contains an analysis of gun and ammunition effects using the Top Ten experiments.

9.9 NBIDE I-2D Analysis for Gun Brand and Ammunition Effects

9.9.1 NBIDE Firing Pin Impressions

Table 9-17 contains average I-2D score values for various matching casing pairs divided into groups by guns and by ammunition combinations. The first three columns of numbers are for the groups where both casings are the same ammunition brand. The columns on the right are when the casings are from different ammunition brands. What are the effects of the brand pairings for gun and ammunition?

Table 9-17. Average I-2D scores from firing pin impressions for casing pairs fired by the same weapon.

Gun ID	Same Ammo Brand				Different Ammo Brands			
	Win.	Rem.	PMC	Average	Win-Rem	Win-PMC	Rem.-PMC	Average
Ruger 41	125	39	136	100	15	102	25	47
Ruger 42	80	54	146	93	71	108	75	85
Ruger 46	66	70	122	86	32	89	22	48
Ruger 48	136	46	144	109	87	127	85	100
Sig Sauer 30	33	38	47	39	26	27	29	27
Sig Sauer 31	29	20	33	27	27	39	24	30
Sig Sauer 32	57	72	62	64	51	48	46	48
Sig Sauer 33	56	65	97	73	23	62	25	37
S&W 305	158	81	184	141	102	159	100	120
S&W 306	69	104	123	99	86	91	107	95
S&W 314	106	120	162	129	99	98	128	108
S&W 401	177	126	152	152	89	115	96	100
Average	91	70	117	93	59	89	64	70

For every gun except Sig Sauer 31, the average score is higher when the reference and compared ammunition brands are the same rather than different. When the ammunition brand is the same for both casings, the average scores are highest when both casings are PMC for every gun except S&W 401 and Sig Sauer 32. The Sig Sauers produce lower match scores, but as will be seen below, they also produce lower non-match scores. The standard deviations of the scores within each group vary widely (from 2 to 40).

Table 9-18 shows the average I-2D scores of casings from non-matching guns of the same brand. The columns are organized by ammunition brand combination. What are the effects of gun brand and ammunition brand?

Table 9-18. Average I-2D scores from firing pin impressions for casing pairs fired by different weapons of the same brand.

Gun Brand	Ammunition					
	Win.	Rem.	PMC	Win-Rem	Win-PMC	Rem.-PMC
Ruger	66	33	73	34	67	32
Sig Sauer	33	26	36	25	31	32
S&W	93	95	114	85	99	94

In a pattern that can seen from the score matrix in Fig. 9-26 and from the histograms in Fig. 9-28, the Sig Sauers produce lower non-match scores along with the lower match scores seen in Table 9-17. In contrast, different guns that are both S&W produce larger scores than for the other two gun brands. When the different guns are both Ruger, there are larger scores in some cases. The standard deviations of the scores within each group vary widely (from 10 to 50).

Table 9-19 shows the mean I-2D scores of casing pairs of the same ammunition brand, but fired from different brands of guns. The standard deviations of the I-2D groups in this table range from 5 to 27. Are there ammunition effects and gun brand interactions?

Table 9-19. Average I-2D scores from firing pin impressions for casing pairs fired by different brands of weapons with the same brand ammunition.

	Win.	Rem.	PMC
Ruger-Sig Sauer	16	16	17
Ruger-S&W	64	42	76
Sig Sauer-S&W	21	20	19

The Rugers and S&Ws correlate much more with each other than with the Sig Sauers, especially when PMC and Winchester ammunition is used for both casings. For the large pool of pairs of casings that have both different gun brands and different ammunition brands, the mean score is 31 and the standard deviation is 26. An analysis of variance (ANOVA) showed that ammunition and gun brand effects, as well as interactions, were significant.

9.9.2 NBIDE Breech Face Impressions

Table 9-20 shows average I-2D score values for various matching casing pairs divided into groups by guns and by ammunition combinations. The first three columns of numbers are for the groups where both casings are the same ammunition brand. The columns on the right are when the casings are from different ammunition brands. What are the effects of the gun and ammunition brands?

Table 9-20. Average I-2D scores from NBIDE breech face impressions for casing pairs fired by the same weapon.

Gun ID	Same Ammo Brand				Different Ammo Brands			
	Win.	Rem.	PMC	Average	Win-Rem	Win-PMC	Rem.-PMC	Average
Ruger 41	81	62	153	99	46	40	28	38
Ruger 42	80	108	185	124	73	101	84	86
Ruger 46	78	170	148	132	42	101	51	65
Ruger 48	89	89	183	120	93	118	102	104
Sig Sauer 30	47	40	63	50	23	38	18	26
Sig Sauer 31	38	42	35	38	25	42	38	35
Sig Sauer 32	71	65	35	57	52	51	35	46
Sig Sauer 33	22	48	37	36	19	24	13	19
S&W 305	72	49	69	63	46	66	43	52
S&W 306	56	61	91	69	48	41	64	51
S&W 314	26	44	75	48	24	33	54	37
S&W 401	66	51	67	61	34	39	38	37
Average	61	69	95	75	44	58	47	50

For every gun, the average score is higher when the reference and compared ammunition brands are the same rather than different. The scores are especially high when both casings are PMC and the gun is a Ruger. The Sig Sauers produce lower match scores, but as will be seen below, they also produce lower non-match scores, although not as much lower as for the firing pin impressions tabulated in the previous subsection. The standard deviations of the scores of those casing pairs from the same gun vary widely from 5 to 67.

Table 9-21 shows the average I-2D score values of non-matching casing pairs fired from different guns of the same brand, organized by ammunition brand combination. What are the effects of gun and ammunition brand?

Table 9-21. Mean I-2D scores from NBIDE breech face impressions for casing pairs fired by different weapons of the same brand.

Gun Brand	Ammunition					
	Win.	Rem.	PMC	Win-Rem	Win-PMC	Rem.-PMC
Ruger	36	51	66	34	41	35
Sig Sauer	19	23	13	17	14	11
S&W	18	24	19	18	15	18

The average non-match scores are highest when both the reference gun and the compared guns are Rugers. The standard deviations of the scores in the group run from 4 to 11 for the Sig Sauers and S&Ws, but from 14 to 40 for the Rugers.

Table 9-22 shows the mean I-2D scores of casing pairs of the same ammunition brand, but fired from different brands of gun. Are there ammunition and gun interaction effects?

Table 9-22. Mean I-2D scores from NBIDE breech face impressions for casing pairs of the same ammunition brand fired by different guns.

	Win.	Rem.	PMC
Ruger-Sig Sauer	11	10	9
Ruger-S&W	17	20	17
Sig Sauer-S&W	11	12	10

These average non-match scores are low, but again the lowest scores occur when one of the guns is a Sig Sauer. The standard deviations of the I-2D scores in this table range from 4 to 9. For the large pool of pairs of casings that have both different gun brands and different ammunition brands, the mean score is 12 and the standard deviation is 7. An analysis of variance showed that ammunition and gun brand effects, as well as interactions, were significant.

There is more evidence of gun and ammunition brand effects and interactions for the I-2D NBIDE data than for the N-3D data of the same casings. The most prominent effects in the I-2D data are:

1. Matching scores are higher if the casings are the same ammunition brand, especially PMC.

2. Sig Sauers give lower scores both for matches and non-matches. The Rugers and S&Ws are better correlated with each other than with the Sig Sauers.

For the N-3D NBIDE data, there are hints of similar trends, but ammunition and gun brand effects appear more pronounced for the I-2D data. Further study is needed for a fuller quantification of these effects.

9.10. Probability Models

The first three subsections of this section discuss some theoretical models and their implications. In Sec. 9.10.4, these models are applied to the N-3D data of the NBIDE casings. It will be clear that all but one of the data sets examined, De Kinder or NBIDE, suggest that current technology is not good enough to support a very large ballistics database. The case of the N-3D analysis of the NBIDE breech faces is a special case and merits a separate discussion, including why it is so different from the De Kinder breech face results.

9.10.1 Simple Binomial Model

In this section we use a binomial model with the overlap metric parameter p to analyze the scenario of a casing from a crime scene being correlated with all the casings in a database. The casings in the database that are chosen for closer scrutiny by a ballistics examiner are those that correlate most highly with the crime scene casing, which will be called the reference casing.

Suppose that there is actually a casing from the same gun in the database, so that it should be a match for the test casing. Let there be N other casings in the database, where N is a suitably large number. For the real match to make a Top Ten list like those produced by the I-2D system, only nine or fewer of the N cross-correlations with non-matching casings may be greater than the cross-correlation with the real match.

For a first pass model, given several simplifying assumptions (Nair [54] has developed a formulation that goes beyond these assumptions), the number of casings in the database that yield a higher cross correlation with the reference casing than does the real match can be modeled as a binomial distribution, Binomial (N, p), where p is the relevant overlap metric. In layman's terms, this is akin to flipping N coins, each with p being the probability of tails, and hoping to get 9 or fewer tails.

In this model, the average number of non-matching correlations higher than the true matching correlation increases linearly with N. De Kinder found empirical evidence of such a linear relationship in his study when investigating the average rank of the true matching casing compared to the other casings.

9.10.2 Some Numbers

This crude probability model makes possible some approximate statements on how good the correlation methods have to be in order to be successful. For instance, suppose that the database has N = 10 000 members with the same class characteristics. How small does p have to be in

order to have the correct casing in the Top Ten at least 99 percent of the time? In probabilistic terms, how small does p have to be in order that, if X is a Binomial (N, p) [55] random variable, the probability,

$$\text{Prob}\,(X < 10) = \sum_{i=0}^{9} \frac{N!}{i!\,(N-i)!} \times p^{i} (1-p)^{N-i} \geq \cdot 0.99\,? \tag{13}$$

One concrete relationship to keep in mind is that by the properties of the Binomial distribution, if $N \times p = 10$, then the probability of the matching casing being in the Top Ten is only around 0.46. Therefore, if N is very large, then p has to be accordingly small. In fact, p needs to be approximately $4/N$ to get the right match in the Top Ten 99 percent of the time. To get in the Top Ten 90 percent of the time, p can be around $6.2/N$.

From this we can make statements of the sort, "If your database is *that* big, then your imaging and correlation techniques better be *that* good to have a reasonable probability of finding a match in it." For instance, if the database has 100 000 entries, then p needs to be on the order of 6.2×10^{-5} or smaller to have a 90 percent chance of getting the correct match in the Top Ten. 100 000 entries has been suggested as a reasonable size for a national database of 9 mm Luger type ammunition [56]. For sake of specificity, this is a typical, representative, and reasonable value for the population size."

9.10.3 Levels of Grouping for Casings and Guns

Note that all of the above applies to the chances for a single casing. Producing a model that describes the performance of a group of casings or a group of guns is more complicated. There are several levels to which the model can be refined. We consider three grouping levels here: a single grouping with a single value of p for all casings and guns, a different grouping for each gun, and a different grouping for each casing. Other types of grouping are also possible, such as grouping by casing brand or by gun brand or by a combination of those.

Single *p*

Suppose the same matching and non-matching distributions can be used for all casings and guns. Then the probability model in the previous sections can be used without modification to refer to all casings such that the same coin with the same p is being flipped for each reference casing. If X is a Binomial (N, p) random variable and we define

$P(N, p)$ = Prob (X <10),

then $P(N, p)$ is essentially the probability that the real match successfully makes it to the Top Ten list, and thus we refer to it as a success rate. Suppose that the specified performance goal is that, given a single matching casing mixed with N non-matching casings in the database, the matching casing is included in a Top Ten list D % of the time. Then, performance is considered satisfactory if $P(N, p) \geq D/100$.

Here we usually set D = 90. The criterion of 90 % seems to be a conservative and reasonable criterion for estimating a desirable efficiency of a large database. This target success rate then

enables us to extrapolate to criteria that might be expected for an experimental database with a small number of entries such as the two collections we studied. The success rate for a small database needs to be extremely high in order to be consistent with the accuracy needs of a large database. More generally, the accuracy criterion depends on whether the gains in the number of matches and the aid to investigators will be sufficient to warrant the cost of developing and maintaining a large database.

Grouped by Guns

There is variability between guns. When there are multiple firings from each gun, we can form separate matching and non-matching distributions for each gun, resulting in a different p for each gun. Thus if a certain gun has an overlap metric p, its casings would tend to make the Top Ten list with probability $P(N, p)$. For a set of guns, each with its own p and $P(N, p)$, then the average success rate of the group of guns is the average of the gun success rates, i.e. the mean of the $P(N,p)$ values for each gun.

To give a simple example, suppose for a fixed database size N there are ten guns of which eight have perfect discrimination ($p = 0$ and $P(N, p) = 1$), and two guns have p so large that $P(N, p) = 0$. Then the average success rate $P(N, p) = 0.8$, so 80 % of the guns' casings would make the Top Ten list.

Another useful success criterion is to consider the proportion of guns that would satisfy a particular success rate for a given N. For the simple example above, suppose that the target success rate is 90 %; then eight of ten would meet the target success rate. For very large N, such as N=100 000, this success criterion will often be close to the average success rate because the individual gun success rates tend to be close to either zero or one for most target success rates.

Grouped by Casing

There can also be variability between casings fired from the same gun. When there are multiple firings from each gun, we can form separate matching and non-matching distributions for each casing, resulting in a different p for each casing. Thus if a certain casing has an overlap metric p, it would tend to make the Top Ten list with probability $P(N, p)$. For a set of casings, each with its own p and $P(N, p)$, then the success rate of the group of casings is the average of the casing success rates, i.e. the mean of the $P(N, p)$ values for each casing. The percentage of casings that satisfy the target success rate for a given N can also be used. Similar to the case for guns, this percentage tends to be close to the average success rate for very large N.

Note the requirement for multiple firings from the same gun. If each gun fired m+1 casings, then each casing has only m correct matches. It may be difficult to get good estimates of p using pair-wise comparisons because of the relatively small number of comparisons that can be made. This can be especially problematic because we are most interested in very small values of p, and the pair-wise comparisons can yield estimates of p only as multiples of $1/(mn)$, where n is the number of non-matching $ACCF_{max}$ values per casing. This may lead to too many estimates of zero for the value of p, as well as estimates of p that may be substantially high, as these estimates depend on how many values from the two samples overlap.

One solution to this problem is to fit continuous distributions to the matching and non-matching samples. These distributions can yield estimates of *p* that are non-zero but very small. Of course there would remain the problem of whether the fitted distribution is an appropriate fit, and how good the fit is, given the limited sample size. In this report, we fit normal distributions using the estimated means and variances of each sample. It is of course possible to use different distributions.

Discussion of Groups

Suppose that the different levels of groups produce substantially different overlap metrics. Then one can draw different conclusions depending on the level of grouping. In general, the more numerous and more refined groups will lead to more optimistic conclusions, while having fewer groups that are more pooled will lead to more pessimistic conclusions. That is because success in a very large database demands a very small value of *p*. Thus, individual casings that have bad distinguishability qualities will increase the estimated *p* of their member group to high levels. Having a smaller group limits the damage done by a single casing. To use a golfing analogy, playing extremely poorly on one hole is much more harmful in stroke play (where every stroke counts) than in match play (where only holes won or lost count). This phenomenon is seen with the NBIDE breech faces.

9.10.4 Experimental Results

Tables 9-23 and 9-24 recap some of the N-3D overlap metric results for the individual casings. Refer to Figs. 9-9, 9-14, 9-19, 9-24, and 9-25 for histograms of the overlap metric estimates.

Table 9-23. Distribution of pair-wise comparison estimates of the overlap metric (*p*) for the individual casing model.

Fraction of estimated *p* measures	=0	≤0.01	≤0.1	Data plotted in
De Kinder FP	0.24	0.31	0.46	Fig. 9-9
De Kinder BF	0.014	0.03	0.21	Fig. 9-14
NBIDE FP	0.18	0.25	0.56	Fig. 9-19
NBIDE BF	0.90	0.95	1.0	Fig. 9-24

Table 9-24. Distribution of the overlap metric (*p*) obtained from normal model estimates for NBIDE BF; data plotted in Fig. 9-25.

Fraction of estimated *p* measures	= 0	$\leq 10^{-6}$	$\leq 10^{-5}$	$\leq 10^{-4}$	$\leq 10^{-3}$	≤0.01	≤0.1
	0.11	0.55	0.63	0.72	0.81	0.94	1

The results are clear-cut for all pair-wise comparison estimates except the NBIDE breech face impressions, which will be discussed at the end of the section. For the other three, let's assume the case of the most optimistic grouping (by casing). For a very large database of size N = 100 000, given the limited resolution of the estimates possible for limited sample sizes, the only estimated values of *p* small enough for a reliable ballistics identification system are those that are essentially zero. From Table 9-23, only 18 % to 24 % of the firing pin impressions satisfy this criterion; the percentage is much lower for the De Kinder breech face impressions. Essentially

the same results occur if N = 10 000. By making use of the relationship described in Sec. 9.10.2, we estimate that even if *N* is as small as 100, the proportion of casings meeting the target success rate of 90 % for Top Ten lists for the De Kinder sets is still less than 50 %. The analogous proportion of casings meeting the target success rate would be less than 56 % for the NBIDE firing pin impressions. For comparing the various technologies applied to different impression sites, the proportion of correct matches found in the Top Ten experiments described earlier can serve as a useful guide to system performance comparisons. These averages are summarized in Table 9-25.

Table 9-25. Proportion of correct matches in Top Ten experiments

I-2D De Kinder FP	3.06/6 = 0.51
I-2D De Kinder BF	1.01/6 = 0.17
N-3D De Kinder FP	3.26/6 = 0.54
N-3D De Kinder BF	2.83/6 = 0.47
I-2D NBIDE FP	3.72/8 = 0.46
I-2D NBIDE BF	5.57/8 = 0.70
N-3D NBIDE FP	5.63/8 = 0.70
N-3D NBIDE BF	7.94/8 = 0.99
I-2D De Kinder BF & FP	3.39/6 = 0.57
N-3D De Kinder BF & FP	4.77/6 = 0.80
N-3D De Kinder BF + FP	4.23/6 = 0.71 (Sum Method)
I-2D NBIDE BF & FP	6.20/8 = 0.78
N-3D NBIDE BF & FP	7.99/8 = 0.999

Again, both I-2D and N-3D did much better on NBIDE Breech Face than on De Kinder Breech Face, and N-3D did very well. The I-2D performance for Firing Pin was similar for the two data sets. For N-3D, the performance for NBIDE Firing Pin was somewhat better than for De Kinder Firing Pin. For the De Kinder set, I-2D was much better and N-3D somewhat better for Firing Pin than for Breech Face. In contrast, for the NBIDE data, both N-3D and I-2D were substantially better on Breech Face than on Firing Pin.

NBIDE Breech Face

The NBIDE Breech Face impressions are drastically different than anything else seen. Under the most optimistic scenario of grouping by individual casings, for a database of size *N*=100 000, 90 % of the casings meet the target success rate, using the pair-wise comparison estimates of *p* (See Table 9-23). For fitted normal model estimates of *p*, Table 9-24 suggests that between 63% and 72% of the casings have *p* metrics of 6.2×10-5 or less, a value of *p* small enough to be consistent with the accuracy requirement suggested in Sec. 9.10.2.

If instead, there is grouping by guns, then only about 50% of the guns have a *p* metric of 6.2×10-5 or less (see Table 9-8). Under the pessimistic scenario of a single group, the estimated mean value of $p = 0.002$ remains over 30 times too large, despite being orders of magnitude smaller than anything else seen.

Discussion

For a technology to be feasible for a very large database, its Top Ten lists should have obtained close to all possible correct matches in a relatively low sample size experiment like those described in this report. Nothing we have seen comes close to achieving such high performance standards except for the N-3D performance on the NBIDE breech face impressions, which suggests that topography methods are a significant advance for breech face analysis. However, any resulting claims for topography images of breech face has to be reconciled with the much less impressive performance of the same technology on the De Kinder casings. Gun differences may be a main cause of the differences, as the De Kinder casings all were fired from Sig Sauers. However, the NBIDE study also included casings fired from Sig Sauers, and the subset of topography results from those Sig Sauer casings are still much better than the topography results for De Kinder Breech Face. It also has been speculated that the higher quality ammunition used in the NBIDE study produced clearer breech face markings. How promising 3D technology is for very large databases depends on whether the NBIDE or De Kinder results are more representative of the challenges faced by a national database. Also, one must remember that only 108 casings were fired by only twelve guns from three different brands. Presumably a larger population of guns would make more likely the presence of unfortunate large correlations from non-matching guns. Also, gun brands and models not covered in the NBIDE study (e.g., low-cost guns that were no longer available as new purchases) may well be more difficult to distinguish and identify than those included.

In order to perform at levels necessary for very large databases, say around 100 000 guns of the same class, the error rates must be very low—so low in fact that for experiments with only 70 or 108 casings, as in this report, essentially the only way to achieve such low error rates is for there to be no overlap between the matching and non-matching samples. While there was considerable separation between matching and non-matching distributions for many of the reference casings, especially those fired from Rugers, others had much less margin for error in that the matching correlations were only slightly larger than the largest of the non-matching correlations. Those matching $ACCF_{\max}$ values would be in danger of being overtaken by non-matching $ACCF_{\max}$ values in a very large database with a much larger population of non-matching $ACCF_{\max}$ values. For each individual casing in the NBIDE set, there were only eight $ACCF_{\max}$ values of casings in the matching sample and 99 $ACCF_{\max}$ values of casings in the non-matching sample. Thus, one can try to estimate the distributions by pooling the matching and non-matching samples for each gun; however, this likely makes the estimated distributions wider than they should be (and in fact would estimate that only half the guns would be successful using the NBIDE breech face impression data). Estimating very low probabilities with moderately low sample sizes continues to be a challenging problem. We used normal probability models for the correlation scores themselves in an attempt to ameliorate the problem. Use of the normal models lowered the success rate for the optimistic scenario of grouping by casing.

The topography methods are of relatively recent vintage, and as such are still being continually refined and improved. There are still questions on the best way of handling data processing and optimal cross correlations, including measures different from the $ACCF_{\max}$. It is possible that refinements will result in great improvements in future. The topography methods look promising for breech face impressions, but improvements are still needed, as is investigation into the factors necessary to obtain the required accuracy.

9.10.5 Normal-Binomial Models

Nair's [54] probabilistic formulation goes beyond the simple binomial model described in Sec. 9.10.2 in dealing with dependencies. In addition, he produces a framing of the problem in terms of modeling the matching and non-matching score distributions by normal models, although he notes that the model does not rely on adherence to normality to produce useful results. Sections 9.8 and 9.9 of this report contain the means and standard deviations of groups of $ACCF_{\max}$ values for various pairings of gun and ammunition types. The standard deviations of the combined groups tend to be larger if the groups are disparate. Nair notes that there will be better database performance if the standard deviations of matching scores are lower than those of non-matching scores (his 'optimistic scenario') and worse if the standard deviations are equal ('pessimistic scenario'). Unfortunately, we have found that in most cases, the matching scores contain more variability (and thus have higher standard deviations) than the non-matching score distributions. Thus, these results using his model would be even less promising than those for his pessimistic scenario. Note for some fingerprint algorithms, it is similarly the case that the non-matching scores are more tightly bunched, and the matching scores are more spread out [53].

10. Statistical Analysis: Gun Distinguishability

10.1 Introduction

The main purpose of the present study is to determine if guns are uniquely distinguishable and identifiable, and if so—for the two data sets at hand (De Kinder and NBIDE)—what that implies about distinguishability/identifiability for a much larger (e.g., national) database of guns.

The starting point for this analysis is the Top Ten list that the N-3D analysis produced. For the De Kinder data, the N-3D cross-correlation analysis produced a Top Ten list for each of the 70 test fired casings. For each of these 70 reference casings, a correlation $ACCF_{max}$ was computed using all of the remaining 69 casings as a comparison group. Since the De Kinder experiment consisted of ten guns and seven ammos per gun, that means that of the 69 comparison casings, 63 (= 9 guns × 7 ammos) will not be a match, while six (= 1 gun × the 6 remaining ammos) will be a match. Hence ideally, a gun would be considered "distinguishable" if all 6 of the remaining ammo firings from that gun show up in the Top Ten list. On the other hand, if only a few (or none) of the six matching casings show up in the Top Ten list for a given reference gun, that particular gun has poor distinguishability. Thus for the De Kinder study, our distinguishability metric is the number of comparative matches (0 to 6) that a particular reference gun ID had in the N-3D Top Ten list. Six represents excellent distinguishability, while zero represents no distinguishability. Further, an individual gun would be considered distinguishable if all seven test firings that used that gun as a reference had all six of their correct matches in the seven Top Ten lists. If all ten guns behaved in this desirable fashion, we would have universal distinguishability for the ten guns (and seven ammos) used in this test.

Similarly, an analogous metric was used for the analysis of the NBIDE data. In this case, however, there were twelve distinct guns, 3 ammos, and 3 repeats (yielding a total of 108 firings) and so a gun would be considered "distinguishable" if for any of its fired casings, all eight (= 3 ammos × 3 repeats – 1) remaining casings fired from that gun appear in the Top Ten list when compared with it. If all twelve of the guns behaved in this fashion, then we would have universal distinguishability for the twelve guns (and three ammos) used in this test.

We now examine the individual gun distinguishability for each of the four combinations of two databases (De Kinder and NBIDE) and two imaging regions (Firing Pin and Breech Face).

10.1.1 Individual Guns (De Kinder / Firing Pin)

The De Kinder experiment had one gun type, ten distinct guns, and seven ammos, for a total of 70 firings. We address the following five questions:

Q1. Are the ten guns distinguishable?
Q2. Are some guns more distinguishable than others?
Q3. What is the best (easiest) gun to distinguish?
Q4. What is the worst (most difficult) gun to distinguish?
Q5. What is the distinguishability ranking (best to worst) of the ten guns?

To address these questions, see the pair of plots in Fig. 10-1. The top plot is a scatter plot; the bottom plot is a mean plot. The horizontal axis of both plots is gun ID. There are ten De Kinder reference gun IDs: 7, 9, 117, 139, 213, 215, 314, 375, 430, and 535. These ten IDs reflect the fact that the De Kinder experiment had ten guns randomly drawn from a larger population of guns.

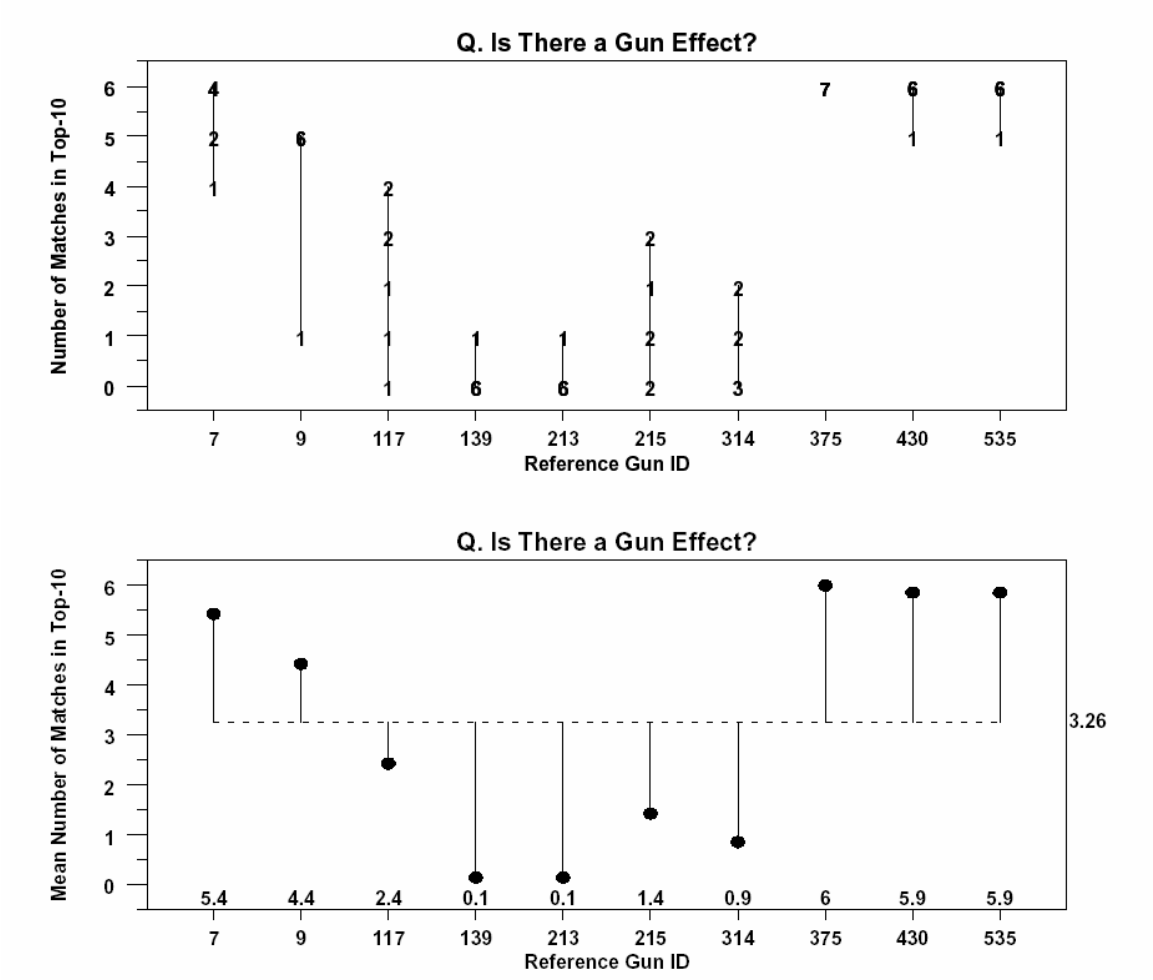

Figure 10-1. De Kinder / Firing Pin $ACCF_{max}$ Top Ten analysis for a gun effect .

The vertical axis of Fig. 10-1 is "score" for an individual gun as a result of the 70 De Kinder firings. Out of those 70 firings, an individual gun is involved in seven such firings. The casings for those seven firings —with each used as a reference with the remaining 69 casings used as a comparison set—result in seven Top Ten lists, and these seven Top Ten lists contain some number of its six matching casings. The vertical axis is 0 to 6. A score of 6 represents perfect distinguishability—for any reference casing, all six remaining casings fired by the same gun appear in the Top Ten list. A score of 0 says that none of the six matching casings appear in the Top Ten list, and hence that reference casing (and gun) is indistinguishable.

Above each reference gun ID on the horizontal axis, there should be 7 marks (one for each time the gun was used in a test firing). To accommodate overstriking, the plot character represents the number of times a score occurs. Hence for the first gun (gun 007), the "4" appearing at the vertical axis value of "6" means that out of the seven firings for this gun, exactly four of the seven Top Ten lists contain all six matches. Moreover for gun "007" there are two Top Ten lists that have five of six matches, and one Top Ten list which has four of six matches.

The ideal distinguishable gun would have "7" at Y = 6 —all seven test firings for this individual gun yielding Top Ten lists having all six of its matching casings. Universal distinguishability would have all ten guns with 7's at Y = 6.

The top plot is the scatter plot of raw scores. The bottom plot is the mean of the seven raw scores for each gun. A reference gun with high (= 6) mean score implies distinguishability; low mean score indicates non-distinguishability. The numbers above the lower horizontal axis (5.4, 4.4, 2.4,...) are the mean scores for each gun.

From Fig. 10-1 we conclude that the guns are not universally distinguishable. Only gun 375 has a perfect score of 6. Six of the ten guns have at least one score of 1 (= poor). The distinguishability ranking of the ten guns is as follows:

1.	375	(mean score = 6.0)	The best gun in terms of distinguishability.
2.	430	(mean score = 5.9)	This gun is near-distinguishable.
3.	535	(mean score = 5.9)	This gun is near-distinguishable.
4.	007	(mean score = 5.4)	Distinguishable for some ammos but not others.
5.	009	(mean score = 4.4)	Distinguishable for some ammos but not others.
6.	117	(mean score = 2.4)	Poor distinguishability for most ammos.
7.	215	(mean score = 1.4)	Poor distinguishability for all ammos.
8.	314	(mean score = 0.9)	Poor distinguishability for all ammos.
9.	139	(mean score = 0.1)	The co-worst. This gun is not distinguishable.
10.	213	(mean score = 0.1)	The co-worst. This gun is not distinguishable.

These De Kinder findings are poor. With ten nominally identical guns, one would expect more consistency across all ten guns. Such consistency is not in evidence. We note finally that the observed differences across the ten guns are statistically significant at the 5 % level (that is, the observed data could happen by chance at most 5 % of the time).

In the ideal, the ten guns should all have perfect scores of 7 at Y=6, and there should not be a statistically significant difference across the ten guns. Neither condition was observed for the De Kinder firing pin impression data, which implies insufficient distinguishability (and hence non-feasibility) for the issue of the much larger national database.

10.1.2 Individual guns (De Kinder / Breech Face)

To address the same five questions as in the previous section, but to focus on the De Kinder breech face data, we make reference to Fig. 10-2. From the figure we conclude that the guns are also not universally distinguishable. At the 5 % level, the differences between guns are statistically significant. Further, the best guns (375, 430, and 535) from the De Kinder firing pin analysis are not the best guns for the De Kinder breech face analysis. The best Breech Face gun (215) was the third worst Firing Pin gun. No Breech Face gun achieved a perfect score of 6, contrary to gun 375's perfect score for Firing Pin.

The average score for the breech face analysis is 2.83 (out of 6), which is smaller than the average score (3.26) for the firing pin analysis, which reaffirms that for the De Kinder data, Firing Pin was a better discriminator than Breech Face.

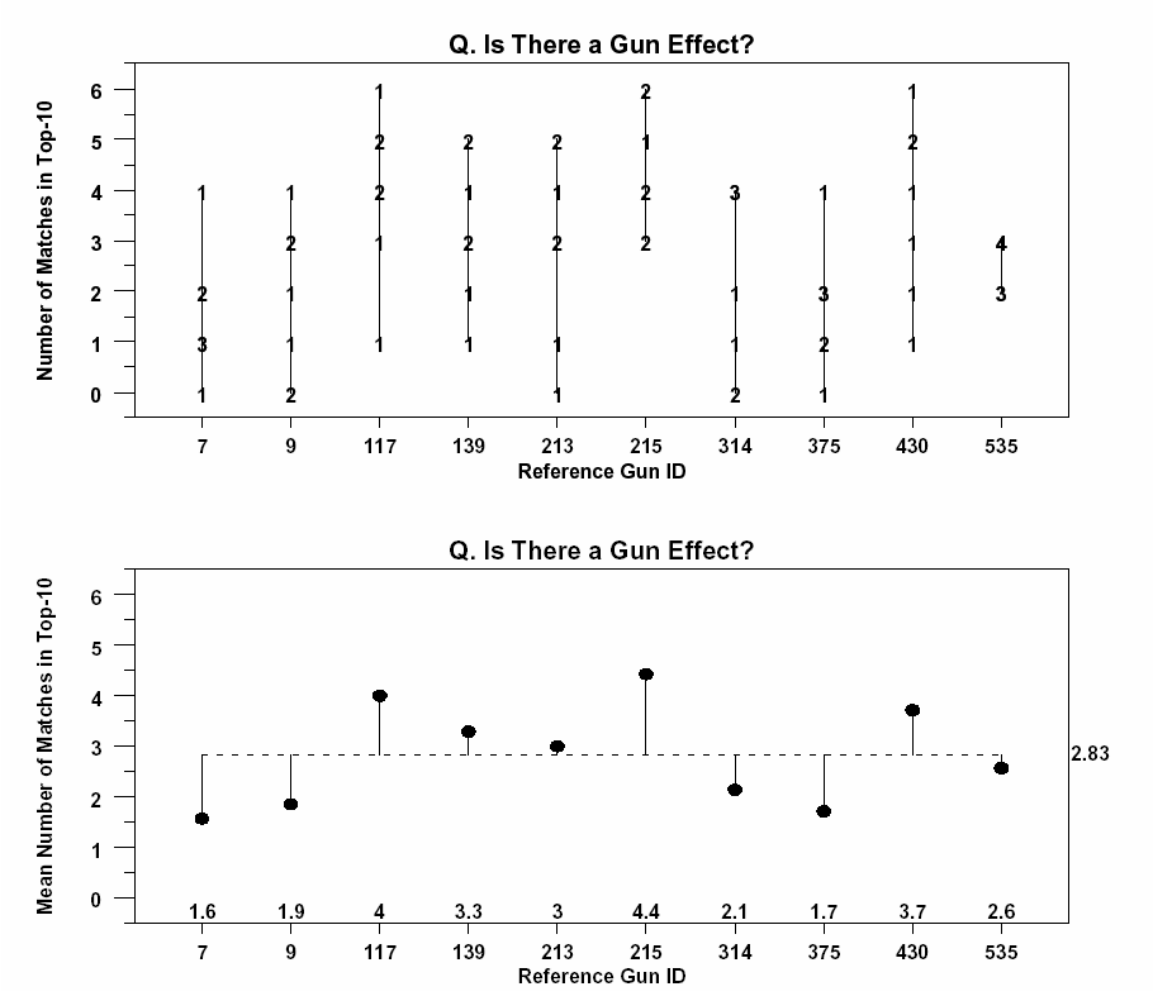

Figure 10-2. De Kinder / Breech Face $ACCF_{max}$ Top Ten analysis for a gun effect.

With 6 being a perfect score, the ranking of the guns for De Kinder / Breech Face is as follows:

1. 215 (mean score = 4.4)

2. 117 (mean score = 4.0)
3. 430 (mean score = 3.7)
4. 139 (mean score = 3.3)
5. 213 (mean score = 3.0)
6. 535 (mean score = 2.6)
7. 314 (mean score = 2.1)
8. 009 (mean score = 1.9)
9. 375 (mean score = 1.7)
10. 007 (mean score = 1.6)

Gun 215 is the most distinguishable gun, with high or moderate scores for all ammos. Guns 117 and 430 are next best, but their scores range from 6 to 1. The remaining guns have only moderate to poor distinguishability. The worst two guns are 007 and 375, which for some ammos had no matches in the Top Ten list.

10.1.3 Individual Guns (NBIDE / Firing Pin)

The NBIDE experiment had three gun types, four replicates of each gun type, a total of twelve distinct guns, three ammos, and three days(reps), for a total of 108 test firings. We here address five questions similar to those of Sec. 10.1.1:

Q1. Are the twelve guns distinguishable?
Q2. Are some guns more distinguishable than others?
Q3. What is the best (easiest) gun to distinguish?
Q4. What is the worst (most difficult) gun to distinguish?
Q5. What is the distinguishability ranking (best to worst) of the twelve guns?

To address these questions, we use Fig. 10-3. The horizontal axis of both plots in Fig. 10-3 shows the twelve NBIDE reference gun IDs: 1, 2, 3, ..., 11, 12. Below these horizontal axis tic labels is a second row of identifying labels: S1, SW5, R1, etc. The labels refer to the gun types as follows:

Gun 1	Sig Sauer 31	S1
Gun 2	S&W 305	SW5
Gun 3	Ruger 41	R1
Gun 4	S&W 306	SW6
Gun 5	Ruger 42	R2
Gun 6	Sig Sauer 32	S2
Gun 7	S&W 401	SW1
Gun 8	Sig Sauer 30	S0
Gun 9	Ruger 46	R6
Gun 10	Sig Sauer 33	S3
Gun 11	S&W 314	SW4
Gun 12	Ruger 48	R8

The last digit of each gun's serial number was used for the digit in the abbreviated ID name (e.g., gun 2 is shortened to SW5 where the 5 comes from the last digit of its serial number: PBV8305). With these identifiers and from Fig. 10-3, we can arrive at initial conclusions about the effect of the three gun types, but we postpone that gun type discussion to a later section (10.2.3), and continue now to focus on the core question at hand of distinguishing between the twelve individual guns.

The vertical axis of Fig. 10-3 shows the results of all 108 firings. There should be a total of 108 data points on the plot (and 9 data points above each reference gun ID), but again due to the over-striking, the plotted character represents the frequency of identical scores. The vertical axis is 0 to 8. A score of 8 represents perfect distinguishability—that is, the chosen reference casing has all eight matching comparison casings (3 ammos × 3 days minus the reference casing) appearing in the Top Ten list. A score of 0 indicates that none of the eight matching casings appeared in the Top Ten list and hence that reference casing/gun is indistinguishable.

As before for the ideal, if a gun were perfectly distinguishable, then there would appear on the plot above the gun ID a "9" at the Y=8 level. If all twelve guns were universally distinguishable, every one of the twelve reference gun IDs would have a "9" at the Y=8 level.

Also as before, the bottom half of Fig. 10-3 is not the nine individual scores, but rather the mean of the nine scores. A reference gun with high mean score implies distinguishability; low mean score indicates non-distinguishability.

From Fig. 10-3 we find that there is a (statistically significant) difference in the twelve guns and the average score is 5.63 (out of 8). Hence we find that the twelve guns are not universally distinguishable. Some guns are more distinguishable than others. Gun 5 has a perfect mean score of 8, followed by gun 2 with a mean score of 7.7. At the other extreme, gun 1 is poor, with none of its nine reference casings having more than five matches in the Top Ten lists. The ranked list of guns is as follows:

1.	5	(mean score = 8.0)
2.	2	(mean score = 7.7)
3.	6	(mean score = 6.9)
4.	3	(mean score = 6.7)
5.	12	(mean score = 6.7)
6.	11	(mean score = 6.4)
7.	10	(mean score = 5.1)
8.	9	(mean score = 4.9)
9.	4	(mean score = 4.3)
10.	7	(mean score = 4.2)
11.	8	(mean score = 4.0)
12.	1	(mean score = 2.7)

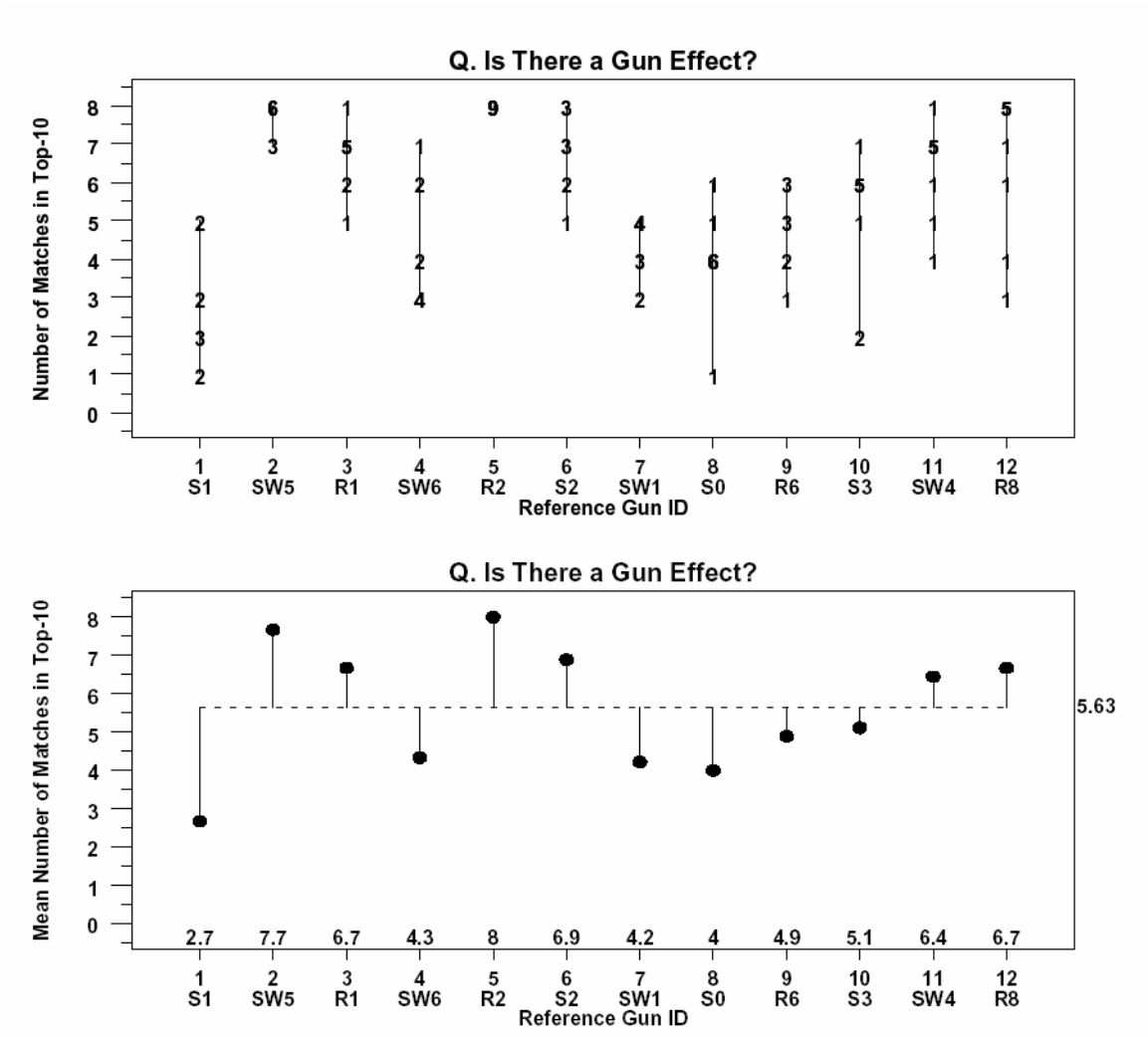

Figure 10-3. NBIDE / Firing Pin $ACCF_{max}$ Top Ten analysis for a gun effect.

10.1.4 Individual Guns (NBIDE / Breech Face)

To address the same five questions as in the previous section, we make reference to Fig. 10-4. From this figure we conclude that the guns for NBIDE / Breech Face are very distinguishable. For seven out of the twelve guns, all nine of their reference test fires had all eight of the correct matching casings show up in the Top Ten lists. All twelve of the guns had at least seven test fires with eight correct matches. The ranked list of guns is as follows:

Rank 1 to 7.	Guns 1, 3, 4, 5, 7, 8, 9	(mean score = 8.0)
Rank 8 to 10 .	Guns 6, 10, 12	(mean score = 7.9)
Rank 11 to 12	Guns 2, 11	(mean score = 7.8)

The average score across all guns is high (7.94 out of 8), and the twelve guns are not statistically different. Hence, of the four data set / image region combinations considered, the NBIDE

Breech Face comes closest to satisfying the visual and statistical requirements for distinguishability.

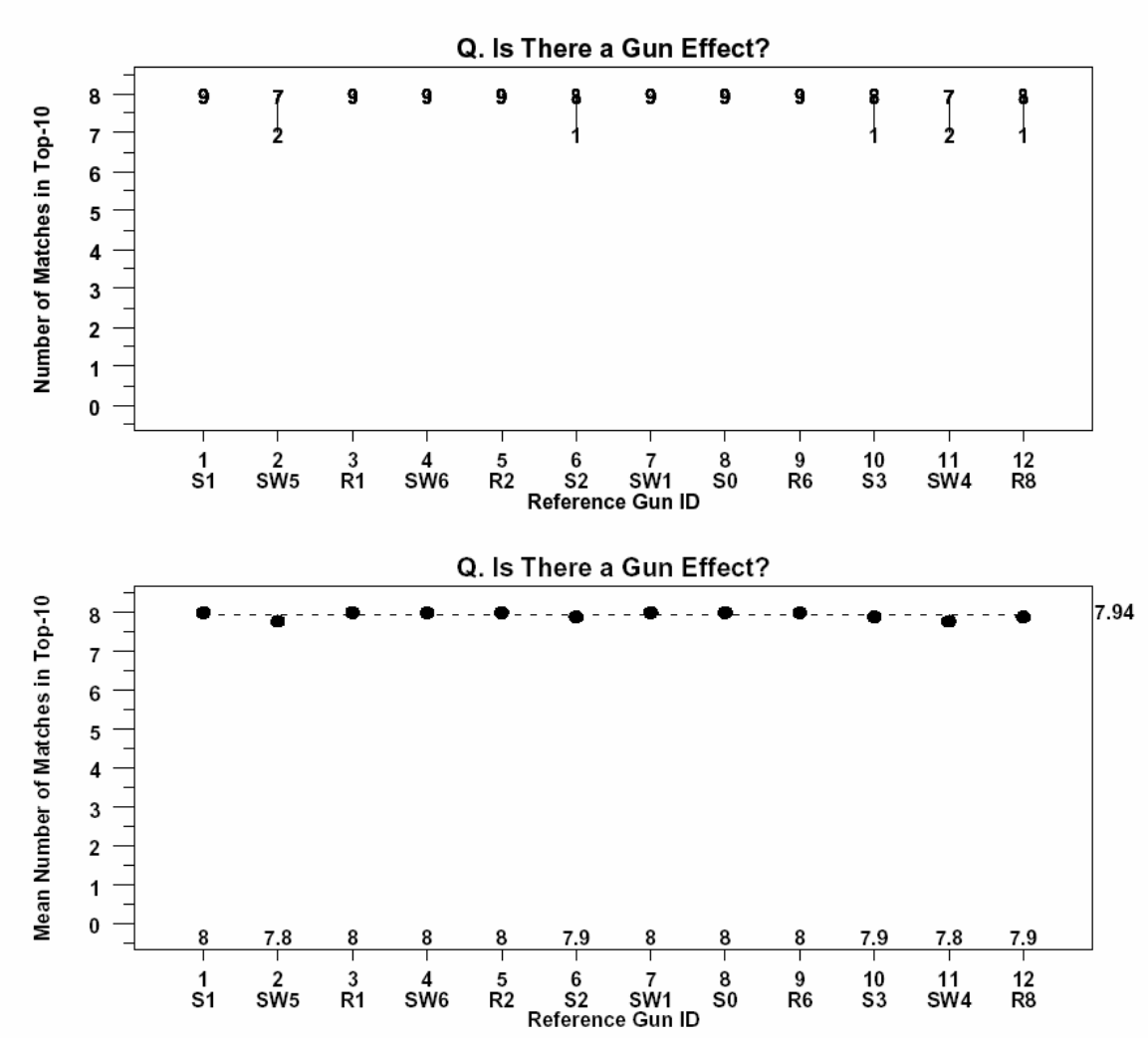

Figure 10-4. NBIDE / Breech Face $ACCF_{max}$ Top Ten analysis for a gun effect.

10.1.5 Individual guns (Summary)

From the above four sections—and especially the mean plots—we find that distinguishability is highly dependent on the source of the data (De Kinder versus NBIDE) and the casing region being imaged (Firing Pin versus Breech Face). Based on the mean number of matches in the Top Ten list, the four cases are ranked as follows (best to worst):

1. NBIDE / Breech Face (mean score = 7.94 out of 8 = 99.3 %)
2. NBIDE / Firing Pin (mean score = 5.63 out of 8 = 70.4 %)
3. De Kinder / FiringPin (mean score = 3.26 out of 6 = 54.3 %)
4. De Kinder / Breech Face (mean score = 2.83 out of 6 = 47.2 %)

As to the core question regarding the feasibility of a (large) national database and whether the imaging and analysis technology is accurate enough to make such a large scale database practical, it is clear that the only one of the four cases that might yield an affirmative is case 4 (NBIDE / Breech Face), with a mean score of 99.3 %. It is of research interest to determine and understand what made this particular combination perform so well. Even at that, major practical hurdles (processing/algorithm speed, gun wear/aging, etc.) would need to be addressed and overcome before the behavior for this small (= 108 firings) data set could be safely extrapolated to a large national system.

10.2 Other Factors Affecting Gun Distinguishability

The prior section deals with the central question of this study, namely whether it is feasible to identify (from a reference casing) the individual gun that fired the casing. We found that the NBIDE Breech Face data provided excellent identifiability, whereas the other cases (NBIDE / Firing Pin, De Kinder / Breech Face, and De Kinder / Firing Pin) yielded unacceptably poorer identifiability.

Given that, a related question arises as to what other factors affect the likelihood of matching a casing with an individual gun. We examine four such factors:

1. Database
2. Imaging Region
3. Gun Type
4. Ammo Type

10.2.1 Database

Database (De Kinder versus NBIDE) is a statistically significant factor. From the discussion in Secs. 10.1.1 through 10.1.4, we find that:

1. The two databases (De Kinder versus NBIDE) are significantly different.

2. The ranking of the two databases (better to worse) is:

NBIDE, mean of 7.94 and 5.63 (= 6.79 out of 8, or 84.9%),
De Kinder, mean of 3.26 and 2.83 (= 3.05 out of 6, or 50.8%).

3. Overall, the best combination is NBIDE / Breech Face.

10.2.2 Imaging Region

Imaging Region is a statistically significant factor. Again, from the discussion in Secs. 10.1.1 through 10.1.4, we find that:

1. The two imaging regions (Firing Pin versus Breech Face) are significantly different.

2. The ranking of the two regions (better to worse) is:
   i. Breech Face (mean of 7.94 out of 8 (99.3%) and 2.83 out of 6 (47.2%) = 73.2% overall),
   ii. Firing Pin (mean of 5.63 out of 8 (70.4%) and 3.26 out of 6 (54.3%) = 62.4% overall).

3. There is an interaction between region and database:
   Firing Pin is better than Breech Face for De Kinder, but
   Breech Face is better than Firing Pin for NBIDE.

4. As before, the best combination is NBIDE / Breech Face.

10.2.3 Gun Type

Note that the question as to whether the matching scores are affected by gun type cannot be addressed from the De Kinder data, inasmuch as all 70 firings came from the same gun type, namely the 9 mm Sig Sauer P226. The NBIDE experiment does, however, shed light on this question with its 108 firings, twelve individual guns, and three gun types:

1. Sig Sauer
2. Smith&Wesson
3. Ruger

Are these three gun types distinguishable? We address this question separately for each of the two imaging regions: Firing Pin and Breech Face.

Gun Type (NBIDE / Firing Pin)
For universal distinguishability, gun type should not have an effect—individual guns should be distinguishable irrespective of gun type. To assess the gun type effect for NBIDE / Firing Pin data, note Fig. 10-5.

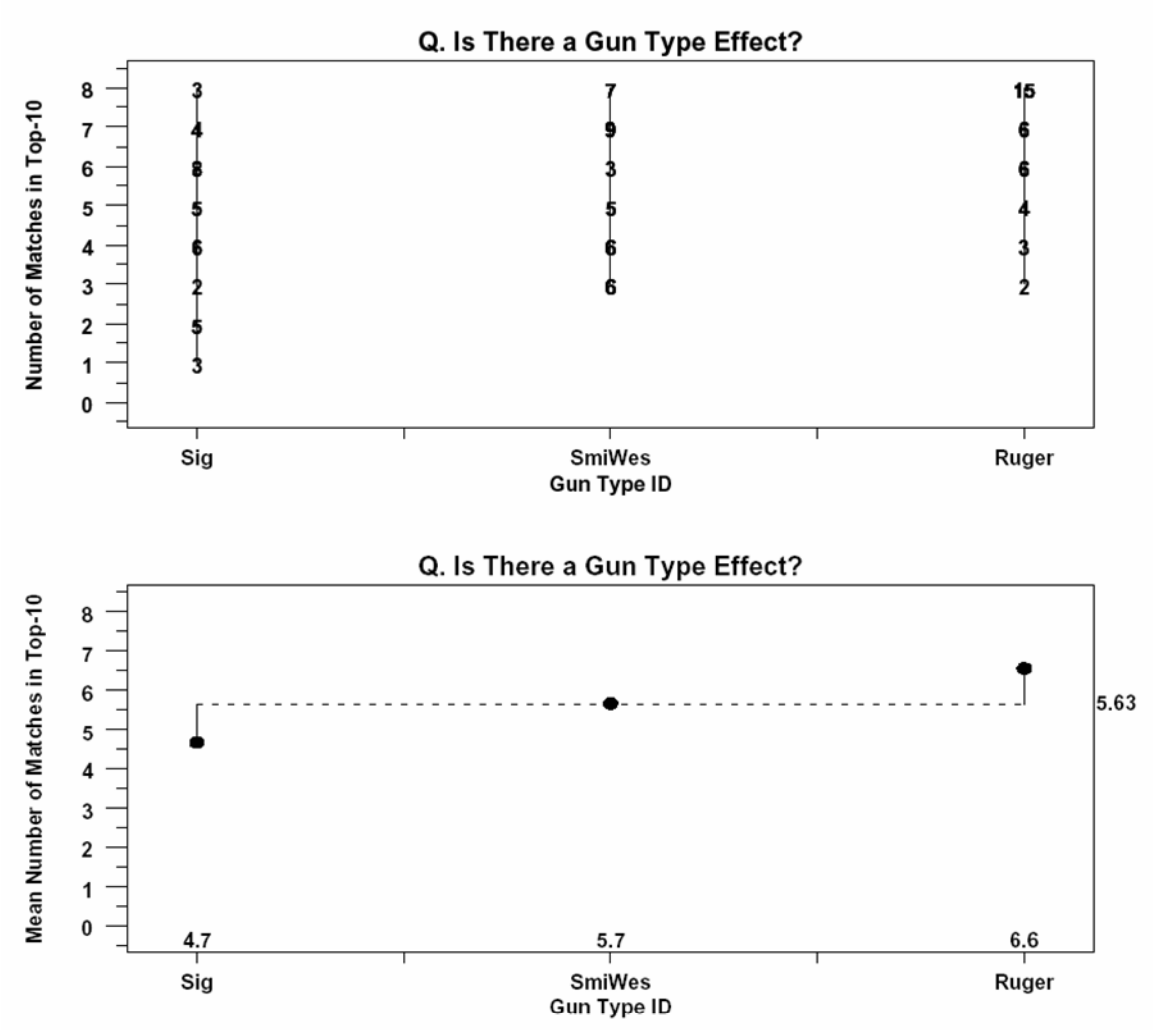

Figure 10-5. NBIDE / Firing Pin $ACCF_{max}$ Top Ten analysis for a gun type effect.

The three gun types are noted on the horizontal axis. As before, the numeric plot character indicates data frequency at that plot point. For example, for Ruger, out of its (108/3 =) 36 test firings, there were fully 15 instances in which all eight of the remaining Ruger casings showed up in the N-3D Top Ten list.

Figure 10-5 indicates that there is in fact a gun type effect—the three gun types are not equivalent in terms of their distinguishability. Some gun types are more amenable to being distinguishable than other gun types. The ranking (best to worst) of the three gun types is as follows:

1. Ruger (mean score = 6.6),
2. Smith&Wesson (mean score = 5.7),
3. Sig Sauer (mean score = 4.7).

The difference across the three gun types is statistically significant at the 5 % level.

Gun Type (NBIDE / Breech Face)
The analysis of gun effects for the NBIDE / Breech face is given in Fig. 10-6.

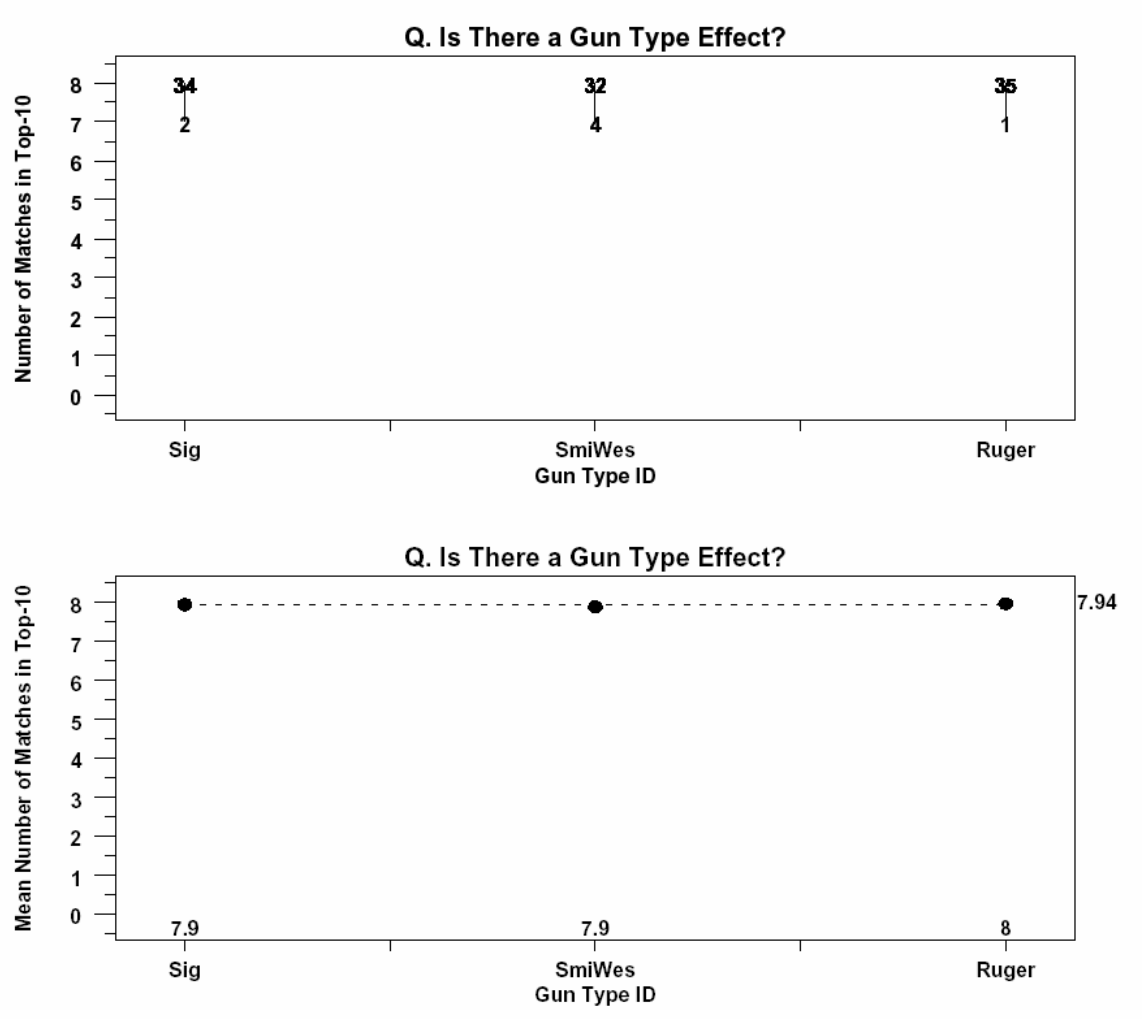

Figure 10-6. NBIDE / Breech Face $ACCF_{max}$ Top Ten Analysis for a gun type effect.

For NBIDE / Breech Face, the three gun types are not statistically significant. This is not unexpected because for this case, almost all of the 108 firings yielded a full score of 8. NBIDE / Breech Face is the ideal: high distinguishability for the twelve guns, with other factors (in particular, gun type) not being statistically significant.

10.2.4 Ammunition Type

Conclusions about the distinguishability of gun type should ideally not be affected by ammunition type. However, it is useful common practice for examiners to use the same brand of ammunition for a testfire as that recovered from a crime scene. Because both the De Kinder and the NBIDE experiments were balanced with respect to ammo (seven ammo types for De Kinder and three ammo types for NBIDE), the analysis for ammo effects is straightforward. We assess the effect of ammo for the usual four cases: De Kinder versus NBIDE, Firing Pin versus Breech Face.

Ammunition Type (De Kinder / Firing Pin)
The De Kinder data set utilized ten guns, all of the same model (Sig Sauer 9 mm Model P226) and seven ammunition (cartridge types):

1. CCI
2. Winchester
3. Remington
4. Speer
5. Wolf
6. Federal
7. Remington (a repeat)

Figure 10-7 examines whether an ammo effect exists. The horizontal axis shows the seven ammo types. The vertical axis is the usual matching score and mean matching score as used in previous figures. For the De Kinder data, each of the seven ammos should have ten points associated with it. If there were no ammo effect, Fig. 10-7 should be near flat, with about the same spread for all ammos. Visually, the ammos are near-equivalent. Quantitatively, at the 5 % level, the seven ammos are not statistically different.

Note that ammo types three and seven are both Remington and thus serve as an internal check on natural variability. As it turns out, the response for the two Remington ammos are near-identical and not statistically different.

Though not statistically different, the ranking of the 7 ammos is as follows:

1.	Speer	(mean score = 3.8)
3.	Federal	(mean score = 3.5)
4.	Wolf	(mean score = 3.4)
5.	Remington	(mean score = 3.2)
6.	CCI	(mean score = 3.1)
7.	Remington2	(mean score = 3.0)
8.	Winchester	(mean score = 2.8)

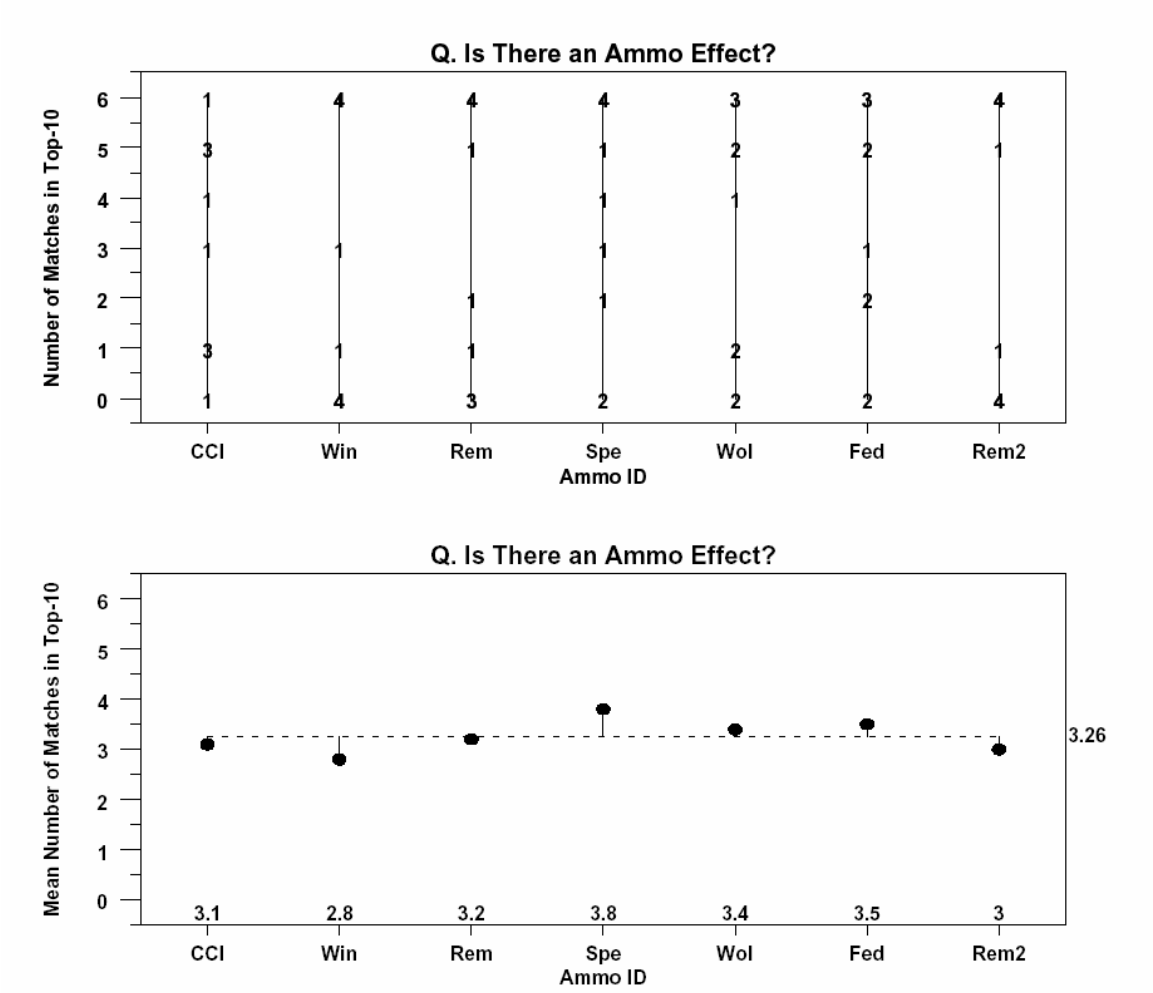

Figure 10-7. De Kinder / Firing Pin $ACCF_{max}$ Top Ten analysis for ammo type.

Ammunition Type (De Kinder / Breech Face)

Figure 10-8 shows the De Kinder / Breech Face data for the question as to whether an ammo effect exists.

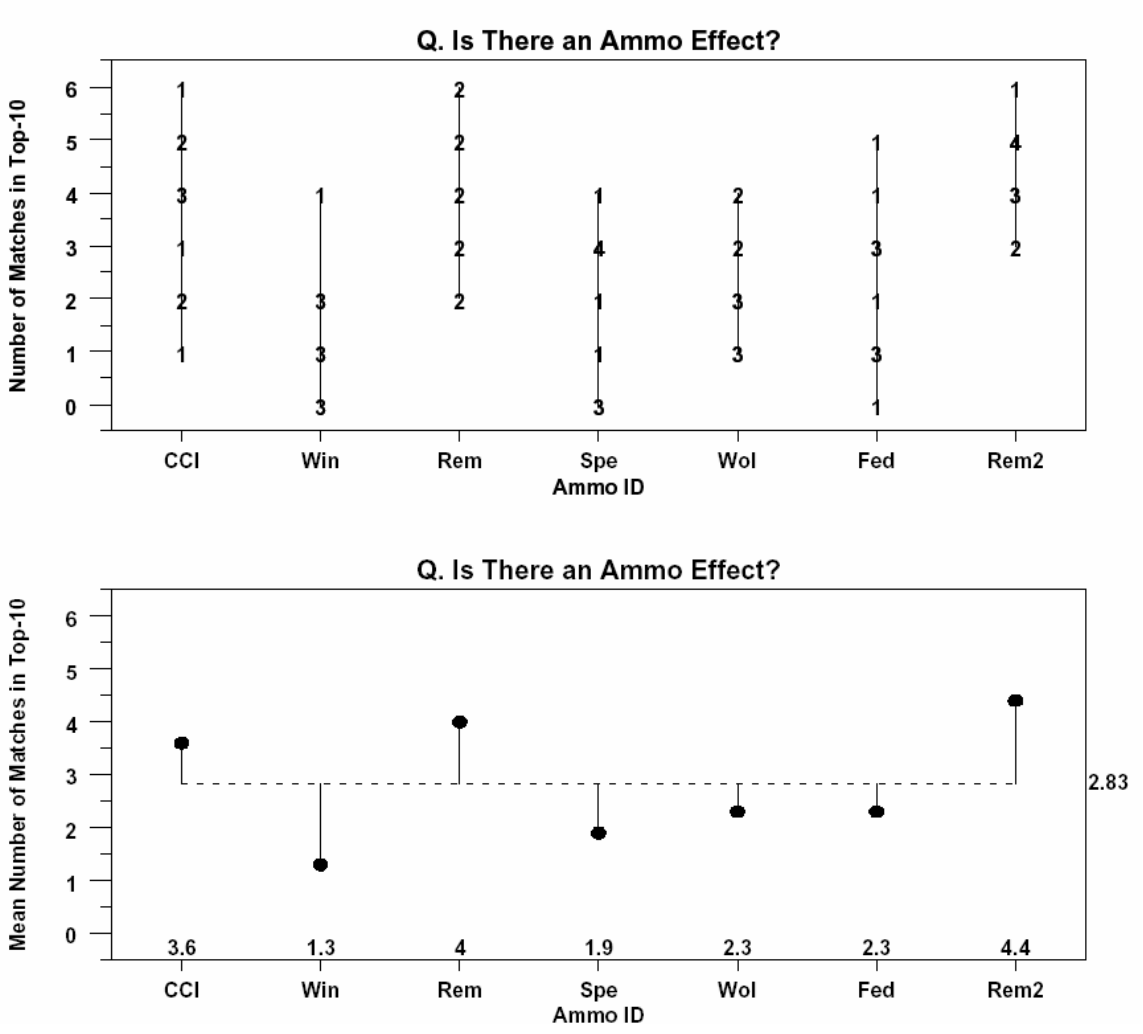

Figure 10-8. De Kinder / Breech Face $ACCF_{max}$ Top Ten analysis for ammo type.

The seven ammos are not equivalent. The seven ammos are statistically different at both the 5 % level and the 1 % level. The two Remington ammos are consistent and higher than the remaining five ammos. The ranking of the seven ammos is as follows:

1. Remington2 (mean score = 4.4)
2. Remington (mean score = 4.0)
3. CCI (mean score = 3.6)
4. Wolf (mean score = 2.3)
5. Federal (mean score = 2.3)
6. Speer (mean score = 1.9)
7. Winchester (mean score = 1.3).

Ammunition Type (NBIDE / Firing Pin)

The NBIDE experiment utilized twelve guns and three ammo types; three days of replication yielded a total of 108 firings. The three ammos were: 1-Remington, 2-Winchester, 3-PMC.

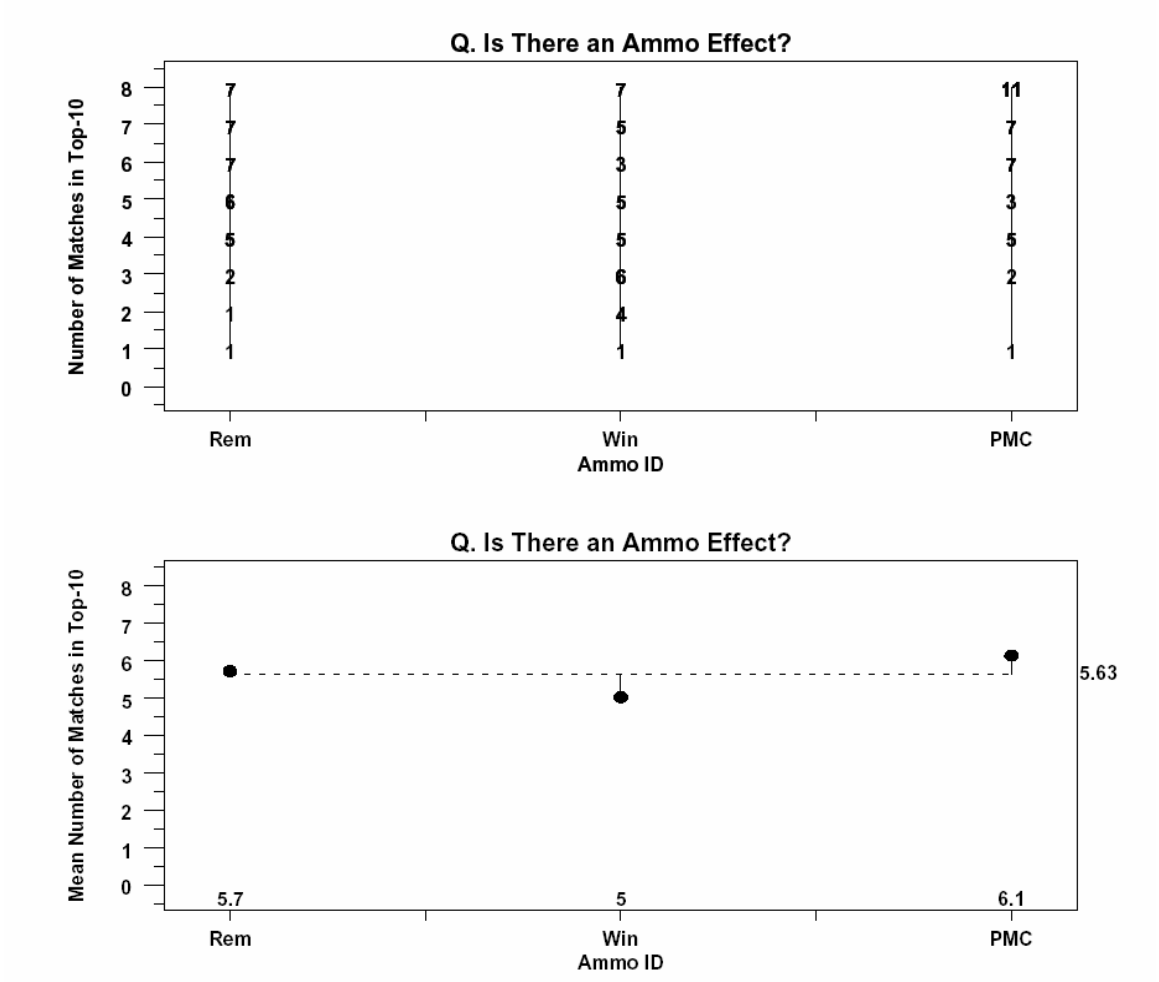

Figure 10-9. NBIDE / Firing Pin $ACCF_{max}$ Top Ten analysis for ammo type.

Figure 10-9 examines whether an ammo effect exists for NBIDE/Firing Pin. Though considerable overlap exists in all three ammos, the graph with PMC having an 11 at Y= 8 suggests that there may be a difference. Statistically, the ANOVA test statistic falls at the 94.5 % point, and so just misses significance at the 5 % level. In short, we reckon ammo type to be marginally significant, with PMC tending to yield more accurate matchings. The ranking of the three ammos is as follows:

1. PMC (mean score = 6.1)
2. Remington (mean score = 5.7)
3. Winchester (mean score = 5.0).

Ammunition Type (NBIDE / Breech Face)

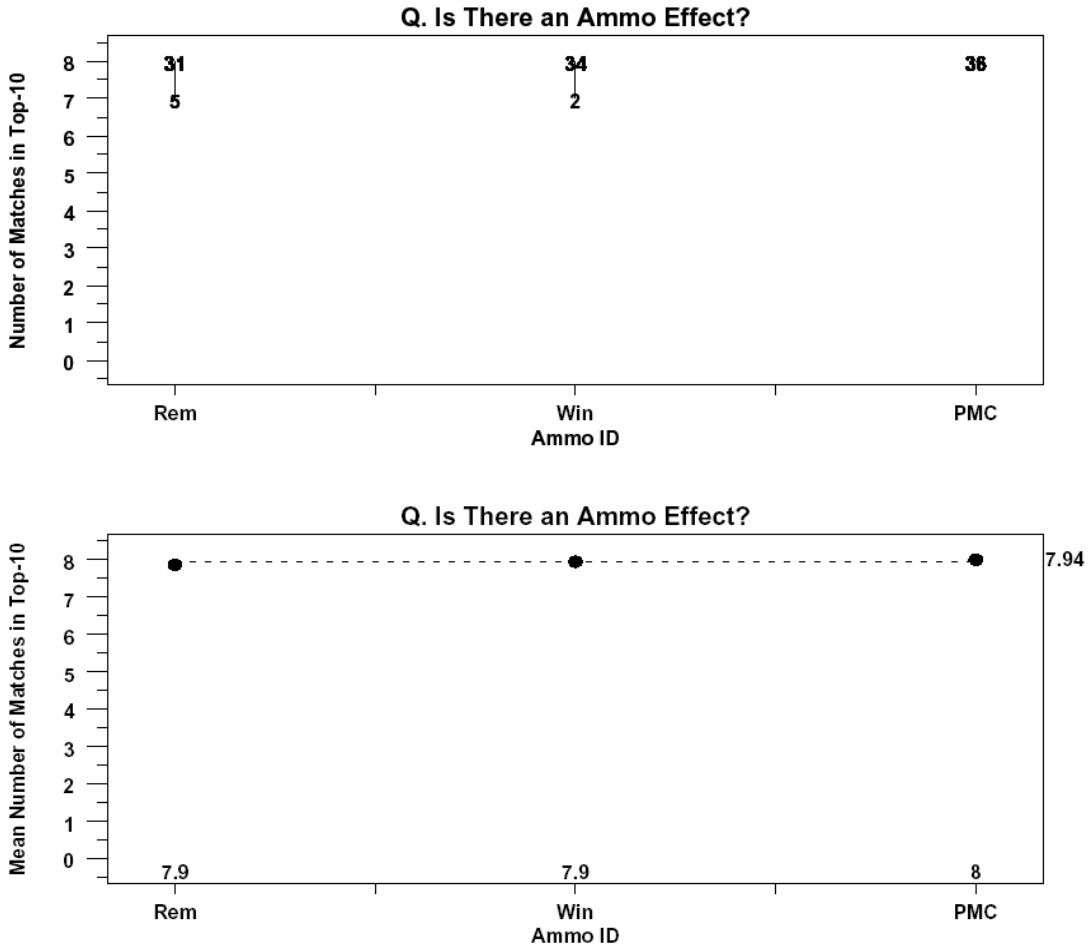

Figure 10-10. NBIDE / Breech Face $ACCF_{max}$ Top Ten analysis for ammo type.

Figure 10-10 investigates for an ammo effect for NBIDE/Breech Face. Because for this case the twelve guns are very distinguishable, the ammo data will necessarily cluster near the *Y*=8 level. There appears to be little difference between the three ammos, although with PMC achieving a perfect score with all of its values at Y=8, there is again the hint that PMC may be doing slightly better than the other two ammos. Statistically, the ANOVA test statistic again falls at the 94.5 % point, and so just misses significance at the 5 % level, which would lead it to being marginally significant. With a maximum difference in the three averages being 0.1, this appears to be a case where the observed differences are statistically significant, but not practically different. The ranking of the three ammos is as follows:

1. PMC (mean score = 8.0)
2. Remington (mean score = 7.9)
3. Winchester (mean score = 7.9)

10.3 Relative Importance of Factors

10.3.1 Relative Importance of Factors (Graphical)

Subsection 10.2 examined individual factors and assessed whether they were significant or not. We finish this section on distinguishability by addressing what is the relative importance of the factors. In particular, we focus on the 3 factors:

1. individual gun
2. gun type
3. ammo

Figures 10-11 through 10-14 examine the relative importance of factors for the usual four cases:

1. De Kinder / Firing Pin` (Fig. 10-11)
2. De Kinder / Breech Face (Fig. 10-12)
3. NBIDE / Firing Pin (Fig. 10-13)
4. NBIDE / Breech Face (Fig. 10-14)

Each individual plot has the multiple factors and the individual factor levels on the horizontal axis, and has the usual mean matching score on the vertical axis. Ideally, for universal distinguishability of gun type, the mean score for the gun should be high, there should be no statistical difference between the individual guns, and there should be no statistical difference among the secondary factors (gun type and ammo).

More to the point, previous analyses have indicated that the various factors are statistically significant in many cases, but here we would like to assess their relative significance. The two De Kinder plots will not have any gun type effect information, since there was only one gun type used (Sig Sauer) in that experiment. From the four plots we conclude:

1. De Kinder / FP: The individual gun effect is more important than the ammo effect.

2. De Kinder / BF: The individual gun effect and the ammo effect are about the same.

3. NBIDE / FP: The individual gun effect is more important than the gun-type effect and the ammo effect, both of which appear to be about the same.

4. NBIDE / BF: The individual gun effect, the gun type effect, and the ammo effect all appear to be negligible. Appearance-wise, this category is markedly different than the other three categories (mean = 7.94).

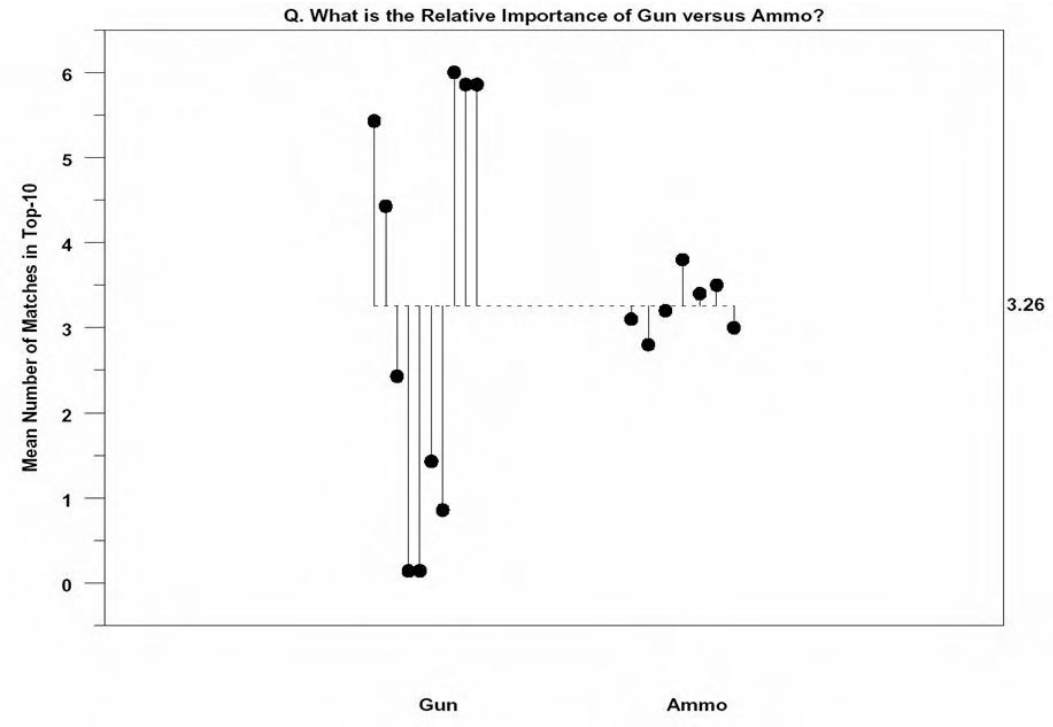

Figure 10-11. De Kinder / Firing Pin $ACCF_{max}$ Top Ten, relative importance of factors.

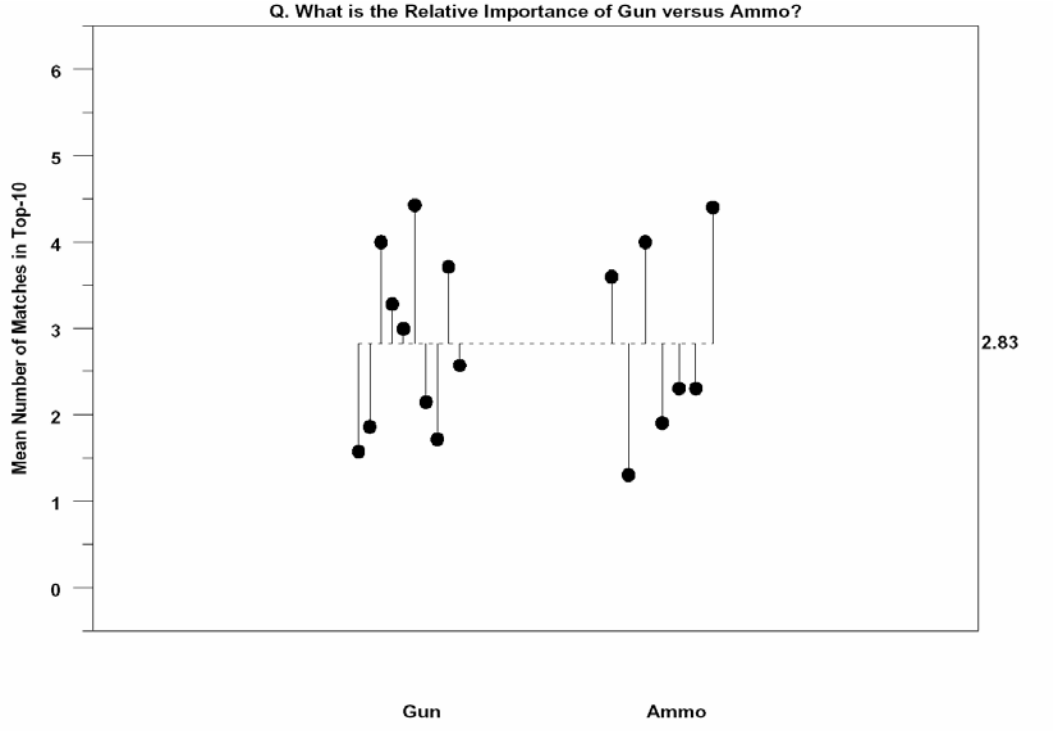

Figure 10-12. De Kinder / Breech Face $ACCF_{max}$ Top Ten, relative importance of factors.

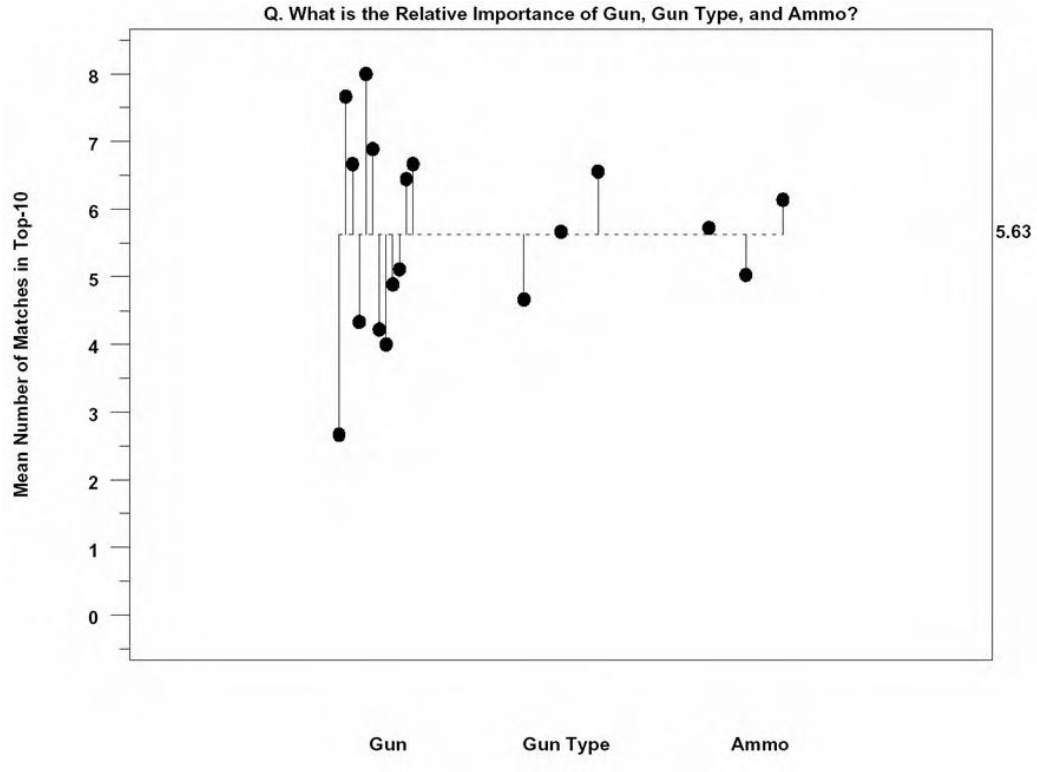

Figure 10-13. NBIDE / Firing Pin $ACCF_{max}$ Top Ten, relative importance of factors.

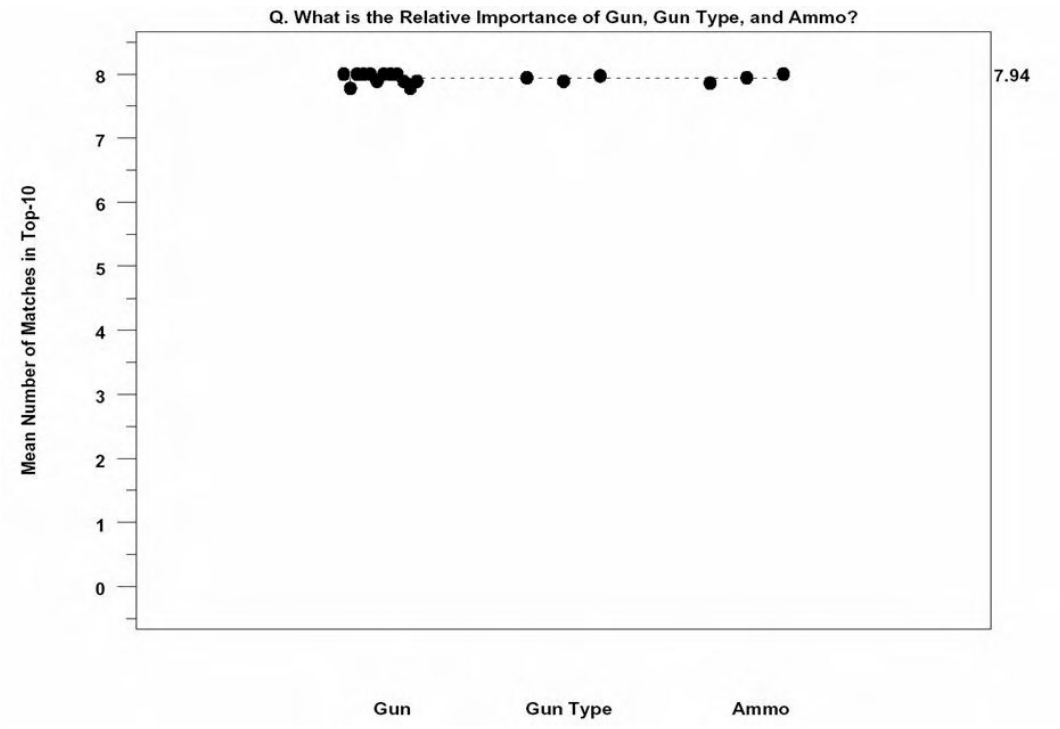

Figure 10-14. NBIDE / Breech Face $ACCF_{max}$ Top Ten, relative importance of factors.

10.3.2 Relative Importance of Factors (ANOVA)

The final subsection deals with examining the relative importance of factors by means of classical (fixed-effect) analysis of variance (ANOVA), which emphasizes the significance testing aspect, in contrast with the random-effects ANOVA, which emphasizes the variance components aspect. We present these ANOVA results for completeness because it allows the assessment of the relative importance of the various factors to be carried out in a formal statistically rigorous complementary fashion. The summary of the ANOVA results is given in Table 10-1 for the four cases graphically presented in Figs. 10-11 through 10-14. The ANOVA does not change the conclusions for the study as a whole. The table is organized as follows:

Column 1: Database and imaging region
Column 2: Summary statistics
Column 3: Results from 1-way ANOVA on gun
Column 4: Results from 1-way ANOVA on gun type
Column 5: Results from 1-way ANOVA on ammo
Column 6: Results from 2-way ANOVA on gun and ammo

From the table we conclude:

1. Gun: Significant at the 1% level for the first 3 cases.
   Not significant for NBIDE / Breech Face.

2. Gun Type: Significant for NBIDE / Firing Pin
   Not significant for NBIDE / Breech Face

3. Ammo: Not significant for De Kinder / Firing Pin
   Significant for De Kinder / Breech Face
   Marginally significant for NBIDE / Firing Pin & Breech Face

Finally, a reminder that “gun” being significant is not good *per se*. In the ideal situation of universal distinguishability, we do not want statistical significance. Rather, we want non-significance in combination with high mean match scores for each and every gun.

Table 10-1. ANOVA Summary Table.
y – the mean of Top Ten Scores.
s – the standard deviation of Top Ten scores.
FCDF – value of the cumulative distribution function for the F–statistics; values ≥ 95% imply significance.
ResSD – residual standard deviation.

Database & Imaging Region	Summary Stat	Gun (k=1)	Gun Type (k=1)	Ammo (k=1)	Gun & Ammo (k=2)
DeKinder Firing Pin	n = 70 Range: 0-6 y = 3.26 s = 2.53	FCDF = 100% ResSD =0.91 Highest Gun: 375 Lowest Guns: 139 & 213 **Significant**	N.A.	FCDF = 1.44% ResSD = 2.63 Highest Ammo: Speer Lowest Ammo: Win Not Significant	1. FCDF = 100% 2. FCDF = 78.27% ResSD =0.89 1. Gun: **Significant** 2. Ammo: Not Sig
DeKinder Breech Face	n = 70 Range: 0-6 y = 2.83 s = 1.63	FCDF = 99.64% ResSD = 1.51 Highest Gun: 215 Lowest Guns: 7 & 375 **Significant**	N.A.	FCDF = 99.98% ResSD = 1.85 Highest Ammo: Rem2 Lowest Ammo: Win **Significant**	1. FCDF = 100% 2. FCDF = 100% ResSD = 1.01 1. Gun: **Significant** 2. Ammo: **Significant**
NBIDE Firing Pin	n = 108 Range: 0-8 y = 5.63 s = 1.99	FCDF = 100% ResSD = 1.26 Highest Gun: Rug2 Lowest Gun: Sig1 **Significant**	FCDF = 99.98% ResSD = 1.85 Highest GT: Rug Lowest GT: Sig **Significant**	FCDF = 94.46% ResSD = 1.95 Highest Ammo: PMC Lowest Ammo: Win Marg. Significant	1. FCDF = 100% 2. FCDF = 99.95% ResSD = 1.17 1. Gun: **Significant** 2. Ammo: **Significant**
NBIDE Breech Face	n = 108 Range: 0-8 y = 7.94 s = 0.25	FCDF = 67.55% ResSD = 0.245 Highest Gun: Many Lowest Guns: SW4&5 Not Significant	FCDF = 65.01% ResSD = 0.247 Highest GT: Rug Lowest GT: S&W Not Significant	FCDF = 94.50% ResSD = 0.243 Highest Ammo: PMC Lowest Ammo: Rem Marg. Significant	1. FCDF = 70.89% 2. FCDF = 94.78% ResSD = 0.240 1. Gun: Not Sig. 2. Ammo: Marg. Sig.

11. Observations and Continuing Work

The topography data and analysis shown here indicate that surface topography measurements may significantly enhance the capability of matching casings fired from the same firearm, particularly for data gathered from breech face impressions. Although the error rate for matching casings was roughly 60 times smaller when topographic data of the NBIDE breech faces were analyzed than when the next most accurate metric was calculated, the error rate would have to decrease by roughly another factor of 35 to adequately support a large database with the assumptions of

- 100 000 guns having the same class characteristics,
- with a single probability distribution of correlation scores,
- and an accuracy goal of 90 %,
- for placing real matches in a Top Ten listing.

Two or three more independent metrics with equal or smaller error rate than that obtained here for matching NBIDE breech face impressions may need to be developed to bring the overall error rate to an acceptable level with the above assumptions.

Segmenting the databases using class characteristics, such as firing pin shape, has been proposed as a way to reduce the sizes of the datasets to be correlated within a large database in order to improve the efficiency of searching a national database for matches. Segmenting by demographic patterns, such as zip code, has also been proposed [57].

The results of the experimental N-3D approach were more accurate than the I-2D results for four experiments. This observation is consistent with results reported by Brinck indicating improved accuracy using IBIS BulletTrax-3D methods to find correct bullet matches as opposed to I-2D methods [58]. Another report by Roberge and Beauchamp reports the successful matching of ten pairs of bullets using IBIS BulletTrax-3D methods [59]. Topography (3D) methods have several advantages:

- Ballistics signatures are mainly geometrical topographies, so a method to measure topography directly should be preferable to reflection microscopy.
- The topography images are not as sensitive to the illumination conditions as reflection microscopy images indicating increased accuracy for 3D methods
- Topography measurements are traceable to dimensional metrology standards.

In addition, the N-3D analysis scheme of outlier removal, filtering, registration, matching, and statistics is non-proprietary, and this openness should facilitate development of improved algorithms by the technical community. For example, standard topography analysis methods [30] may be adapted to separate micro- from macro-topography and extract individual characteristics of the surfaces for correlation and identification.

A disadvantage of the current prototype topography approach is the time required to record the data, which is considerably longer for casings than that required for I-2D. The data gathering procedure and likely the analysis algorithms would need to become much more efficient for

practical application to a large database. The system described here is experimental and is not intended for commercialization.

A second metric might come from using the topography of the ejector marks. During this study we have not used the information from the ejector marks. In particular, we have not developed a technique for correlating different ejector marks because of their widely varying outer boundaries. It is difficult to develop automated software to correlate the shapes of such regions, particularly when some ejector marks are partially obliterated by labels imprinted on the casing by the manufacturer. Common practice for the I-2D is a manual operation whereby the users draw the ejector mark boundaries themselves when making entries. One of our tasks for future work is to develop a similar analysis program for the existing ejector mark data.

A second task not yet completed here is the correlation analysis for the 176 IAI bullets we are measuring. The correlation results could then be compared both with I-2D correlations from image acquisitions at ATF by Ols and Simmers, and with topography images previously measured by Bachrach et al. [15,35] using a single point confocal system. We will also be able to explore whether a sufficiently reliable metric can be developed for bullets to be consistent with a large database of bullet entries.

12. Acknowledgements

We are grateful to D. Cork of the National Academies and J. Rolph and V. Nair of the National Academies' Panel for their valuable guidance during the course of this work and to W. Eddy also of the Panel for valuable discussions. We are also grateful to M. Ols of the ATF for key discussions related to firearms and ballistics selection, to R. Simmers of the ATF for expertise with I-2D acquisitions, and to D. Xiang of IAI for topography data acquisition and analysis. We also thank M. McLean of Forensic Technology Incorporated for providing special data analysis of I-2D acquisitions and correlations. N. Waters of NIST kindly assisted in the NBIDE test firing procedure. The cover was designed by B. Young. Thanks also go to K. Rice and R. Rhorer for their careful reading of the manuscript. The project was supported by the Department of Justice under National Institute of Justice Grant Number 2003-IJ-R-029 with the NIST OLES.

13. References

1. P.L. Kirk, *Crime Investigation*, 2nd Ed. (Krieger Publ. Co., Malabar FL, 1985) p. 364.
2. A.A. Braga and G.L. Pierce, Linking Crime Guns: The Impact of Ballistics Imaging Technology on the Productivity of the Boston Police Department's Ballistics Unit, Paper ID JFS2003205, *J. Forensic Sci.* **49**, 1 (2004).
3. http://www.forensictechnologyinc.com/p7.html, IBIS Heritage Systems (2006)
4. W.C. Boesman and W.J. Krouse, *CRS Report to the Congress, National Integrated Ballistics Information Network (NIBIN) for Law Enforcement*, Order Code RL31040 (Congressional Review Service, Library of Congress, Washington DC, 2001).
5. http://www.nibin.gov/nibin.pdf, ATF's NIBIN Program, June 2005.
6. F.A. Tulleners, Attachment A, Technical Evaluation: Feasibility of a Ballistics Imaging Database for all New Handgun Sales in B. Lockyer, *Feasibility of a California Ballistics*

Identification System, Assembly Bill 1717 (Hertzberg) (Stats. 2000, ch. 271) Report to the Legislature (California Department of Justice, Sacramento CA, 2003).
7. J. De Kinder, F. Tulleners, and H. Thiebaut, Reference Ballistic Imaging Database Performance, *Forensic Sci. Int.* **140**, 207 (2004).
8. W. George, A Validation of the Brasscatcher Portion of the NIBIN/IBIS System, *AFTE J.* **36**, 286 (2004) and A Validation of the Brasscatcher Portion of the NIBIN/IBIS System Part Two: "Fingerprinting Firearms" Reality or Fantasy, *AFTE J*. **36**, 289 (2004).
9. A. Beauchamp and D. Roberge, Model of the Behavior of the IBIS Correlation Scores in a Large Database of Cartridge Cases, http://www.forensictechnology.com/d4.html, accessed 16 April 2007.
10. R. Nennstiel and J. Rahm, A Parameter Study Regarding the IBIS Correlator, *J. Forensic Sci.* **51**, 18 (2006) and An Experience Report Regarding the Performance of the IBIS Correlator, *J. Forensic Sci.* **51**, 24 (2006).
11. http://www.ncjrs.org/txtfiles/165476.txt, J. Travis, Guns in America: National Survey on Private Ownership and Use of Firearms, NIJ Research in Brief (National Institute of Justice, Washington DC, May 1997) p. 12.
12. http://www.troopers.state.ny.us/Firearms/CoBIS/ , New York State Division of State Police, Combined Ballistic Identification System, Accessed, 06 Dec 2005.
13. J.J. Tobin, Jr. *Maryland- IBIS Integrated Ballistics Identification System* (Maryland State Police Forensic Sciences Division, Pikesville MD, 2003).
14. B. Lockyer, *Feasibility of a California Ballistics Identification System, Assembly Bill 1717 (Hertzberg) (Stats. 2000, ch. 271) Report to the Legislature* (California Department of Justice, Sacramento CA, 2003), including Attachments A (Ref. 6), B, C, and D.
15. B. Bachrach, A Statistical Validation of the Individuality of Guns Using 3D Images of Bullets, http://www.ncjrs.gov/pdffiles1/nij/grants/213674.pdf, March 2006.
16. J. Song, E. Whitenton, D. Kelley, R. Clary, L. Ma, S. Ballou, and M. Ols, SRM 2460/2461 Standard Bullets and Casings Project, *J. Res. Natl. Inst. Stand. Technol.* **109**, 533 (2004).
17. L. Ma, J. Song, E. Whitenton, A. Zheng, T. Vorburger, and J. Zhou, NIST Bullet Signature Measurement System for RM (Reference Material) 8240 Standard Bullets, *J. Forensic Sci.* **49,** 649 (2004).
18. A. Harasaki, J. Schmit, and J. C. Wyant, Improved Vertical Scanning Interferometry, *Appl. Opt.* **39**, 2107 (2000).
19. M.A. Schmidt and R.D. Compton, Confocal Microscopy, in *ASM Handbook Volume 18 Friction, Lubrication, and Wear Technology,* P.J. Blau, ed., (ASM International, 1992), p. 357.
20. T.R. Thomas, ed., *Rough Surfaces* (Longman, Harlow, UK, 1982), Chap. 2.
21. J.M. Utts and R.F. Heckard, *Statistical Ideas and Methods* (Thomson Brooks/Cole, Belmont CA, 2006).
22. J.M. Bennett and L. Mattsson, *Introduction to Surface Roughness and Scattering* (Optical Society of America, Washington DC, 1989).
23. O.C. Wells, *Scanning Electron Microscopy* (McGraw-Hill, New York, 1974).
24. J. Song, T. Vorburger, T. Renegar, H. Rhee, A. Zheng, L. Ma, J. Libert, S. Ballou, B. Bachrach and K. Bogart, Correlation of Topography Measurements of NIST SRM 2460 Standard Bullets by Four Techniques, *Meas. Sci. and Technol.* **17**, 500 (2006).

25. M. Bray, Stitching Interferometry and Absolute Surface Shape Metrology: Similarities, *Proc. SPIE*. **4451** (2001); http://www.mboptique.com/common/publications/MBO_2001_SPIE_4451-40_(Stitching_Interferometry).pdf.
26. J. Song and T. Vorburger, Proposed Bullet Signature Comparisons Using Autocorrelation Functions, *Proc. 2000 Nat. Conf. Standards Laboratories* (Toronto, July 2000).
27. http://www.fti-ibis.com/en/s_4_1_5.asp, Bulletrax-3D, Accessed, 06 Dec. 2005.
28. H.-G. Rhee, T.V. Vorburger, J.W. Lee, and J. Fu, Discrepancies between Roughness Measurements Obtained with Phase-Shifting and White-Light Interferometry, *Appl. Opt.* **44**, 5919 (2005).
29. T.V. Vorburger, H.-G. Rhee, T.B. Renegar, J.-F. Song, and A. Zheng, Comparison of Optical and Stylus Methods for Measurement of Surface Texture, *Int. J. Adv. Manuf. Technol*. DOI 10.1007/s00170-007-0953-8, http://www.springerlink.com/content/0851313276m3t772/, online February 2007 (in press).
30. ASME B46.1-2002, *Surface Texture (Surface Roughness, Waviness, and Lay)* (Amer. Soc. Mech. Engrs., New York, 2003).
31. J.F. Song, T.V. Vorburger, R. Clary, E. Whitenton, L. Ma, and S. Ballou, Standards for Bullets and Casings, *Materials Today* **5** (11), 26 (2002).
32. P. Rubert, Properties of Electroformed Calibration Standards for Surface Topography Measurement Systems, in *Tenth International Colloquium on Surfaces*, edited by M. Dietzsch and H. Trumpold (Shaker-Verlag, Aachen, 2000) p. 245.
33. A. Moenssens et al., *Scientific Evidence in Civil and Criminal Cases*, 4th Edition (the Foundation Press Inc., 1995, New York) p. 375.
34. ISO 25178-6, Committee Draft, *Geometrical product specification (GPS)—Surface texture: Areal— Part 6: Classification of methods for measuring surface texture* (International Organization for Standardization, Geneva, 2005).
35. B. Bachrach, Development of a 3D-based Automated Firearms Evidence Comparison System, *Journal of Forensic Sciences* **47**, 1253 (2002).
36. J.H. Dillon, Three Dimensions: The Next Level for Firearms Examination, *Evidence Technology Magazine*, July-August, 34 (2005), www.EvidenceMagazine.com.
37. *Crime Gun Trace Reports (2000) National Report* (Department of the Treasury, Bureau of Alcohol, Tobacco, and Firearms, Washington DC, 2002).
38. µsurf, Non contact 3D-measurement of complex surfaces, http://www.nanofocus.info/product_overview.php?productId=7, Dec 2005, accessed 06 Dec. 2005.
39. Infinite Focus,http://www.alicona.com/.
40. R. Simmers and M. Ols, private communication.
41. Y. B. Yuan, T.V. Vorburger, J. F. Song, T. B. Renegar, A Simplified Realization for the Gaussian Filter in Surface Metrology, in *X. International Colloquium on Surfaces*, M. Dietzsch, H. Trumpold, eds. (Shaker Verlag GmbH, Aachen, 2000), p. 133.
42. Bergen, J. R., Anandan, P., Hanna, K., and Hingorani, R., Hierarchical model-based Motion Estimation, in *Proceedings of Second European Conference on Computer Vision*, (Springer-Verlag, 1992), pp. 237-252.
43. Heeger, D., Notes on motion estimation, prepared by Prof. David J. Heeger for Courses Psych 267/CS 348D/EE 365, New York University, 20 Oct. 1996.
44. Heeger, D., MATLAB™ Software for Image Registration, Copyright 1997, 2000 by Stanford University, available at URL http://www.cns.nyu.edu/~david/registration.html.

45. G.M. Jenkins and D.G. Watts, *Spectral Analysis and Its Applications*, (Holden-Day, San Francisco, 1968) p. 171 ff.
46. *Guide to the expression of uncertainty in measurement* (GUM) (International Organization for Standardization, Geneva, 1995).
47. B. N. Taylor and C. E. Kuyatt, *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, NIST Tech. Note 1297, (1994).
48. Latin Square and Related Designs, in *Engineering Statistics Handbook*, Sec. 5.3.3.2.1, http://www.itl.nist.gov/ div898/handbook/, updated 18 July 2006.
49. R.J. Grissom, Probability of the superior outcome of one treatment over another, *Journal of Applied Psychology* **79,** 314 (1994).
50. D.M. Green and J. Swets, *Signal Detection Theory and Psychophysics* (John Wiley, New York, 1966).
51. J.A. Hanley and B.J. McNeil, The Meaning and Use of the Area under an ROC Curve, *Radiology*, **143**, 129 (1982).
52. H.B. Mann and D.R. Whitney, On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other, *Annals of Mathematical Statistics* **18,** 50 (1947).
53. L. Hong and A.K. Jain, Integrating Faces and Fingerprints for Personal Identification, *IEEE Transactions PAMI* **20**, 1295 (1998).
54. V. Nair, private communication.
55. R.S. Burington, *Handbook of Mathematical Tables and Formulas* (McGraw-Hill, New York, 1965) p.357.
56. F. A. Tulleners, Ref. 6, Secs. 1.5 and 4.5.
57. Forensic Technology, Inc., Segmenting Tool Mark Image Reference Files (TMIRF) to Identify Crime Guns More Effectively, http://www.forensictechnology.com/d4.html, accessed 16 April 2007
58. T. Brinck, Comparing the Performance of IBIS and Bullet TRAX-3D Technology Using Bullets from Ten Consecutively Rifled Barrels, AAFS Meeting, San Antonio, February 2007, http://www.forensictechnologyinc.com/d4.html, accessed 16 April 2007.
59. D. Roberge and A. Beauchamp, The Use of BulletTrax-3D in a Study of Consecutively Manufactured Barrels, *AFTE Journal* **38**, 166 (2006).

JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, March 2010, Vol. 55, No. 2
doi: 10.1111/j.1556 4029.2009.01221.x
Available online at: interscience.wiley.com

PAPER

CRIMINALISTICS

Benjamin Bachrach,[1] *Ph.D.; Anurag Jain,*[1] *M.S.; Sung Jung,*[1] *M.S.; and Robert D. Koons,*[2] *Ph.D.*

A Statistical Validation of the Individuality and Repeatability of Striated Tool Marks: Screwdrivers and Tongue and Groove Pliers*

ABSTRACT: Tool mark identification relies on the premise that microscopic imperfections on a tool's working surface are sufficiently unique and faithfully transferred to enable a one to one association between a tool and the tool marks it creates. This paper presents a study undertaken to assess the validity of this premise. As part of this study sets of striated tool marks were created under different conditions and on different media. The topography of these tool marks was acquired and the degree of similarity between them was quantified using well defined metrics. An analysis of the resulting matching and nonmatching similarity distributions shows nearly error free identification under most conditions. These results provide substantial support for the validity of the premise of tool mark identification. Because the approach taken in this study relies on a quantifiable similar ity metric, the results have greater repeatability and objectivity than those obtained using less precise measures of similarity.

KEYWORDS: forensic science, tool mark identification, 3D imaging, automated comparison of microscopic tool mark evidence, striations, screwdrivers, tongue and groove pliers, statistical methodology

The ability to perform tool mark to tool mark comparisons based on microscopic features observed on the tool mark's surface is at the core of tool mark identification. Supreme Court decisions such as *Daubert versus Merrill Dow* (1) and *Kumho Tire versus Carmi chael* (2) are making it increasingly necessary to further formalize scientific evidence presented in court. Furthermore, the develop ment of DNA identification techniques and the level of accuracy achievable in the estimation of the associated error rates have raised the expectations for the quantitative precision that may be achieved in forensic analysis. Quantitative evidence regarding the validity of the basic premise of tool mark comparison would pro vide additional support for the admissibility of tool mark evidence. The Federal Bureau of Investigation (FBI) and Intelligent Automa tion, Inc. (IAI) have undertaken an extensive study to verify the premise that the microscopic features transferred from a tool's working surface to the marks created by it are sufficiently unique and repeatable to enable the association of a tool with its marks. This paper reports the results of this study for the case of striated tool marks (a paper reporting the results for impressed tool marks is in preparation). In particular, we consider two types of tools: screwdrivers and tongue and groove pliers. In addition to consider ing the comparison of striated tool marks created under the same conditions, we also evaluated the effect of the media onto which the tool marks are created. In the case of screwdrivers, the effect of the variation of angle of attack in the creation of striated tool marks was also evaluated.

An important element of this study was the use of topographical (3D) data for the characterization of tool marks. The concept of using a 3D characterization of a surface for identification purposes goes as far back as 1958, when Davis (3) proposed the idea of the "Striagraph" for ballistic identification. The application of 3D methodologies to obtain characteristic information about striated marks on bullets has also been reported by DeKinder (4,5). Geradts (6) has presented a system capable of performing, in an automated way, comparisons between 3D topographical measurements of striated tool marks. Bachrach (7) has described an automated com parison system that uses 3D information of a bullet's surface to perform automated comparisons. More recently, Banno (8) has reported on the 3D visualization and comparison of features on fired bullets by using 3D surface topography data. The principles of tool mark identification can be found in Miller (9). An often cited study on the criteria for identification for firearm and tool mark identification was published by Biasotti and Murdoch (10). Another significant effort that examines the theory of identification as it pertains to tool marks and the criteria for their identification has been reported by Miller (11). An exhaustive review of the liter ature pertaining to the identification criteria for firearms and tool mark identification has been more recently carried out by Nichols in 1997 (12) and 2003 (13). This study builds upon and extends the results of the previous studies by providing consistent quantita tive measures in 3D of tool mark similarity.

As part of the study reported in this paper, a confocal micro scope was used to acquire topographical data of tool mark samples. A significant number of striated tool mark samples were created under controlled conditions on a variety of media. Algorithms were

[1]Intelligent Automation, Inc., 15400 Calhoun Drive, Suite 400, Rockville, MD 20855.
[2]Counterterrorism and Forensic Science Research Unit, FBI Laboratory, FBI Academy, Quantico, VA 22135.
*Presented in part at the Association of Firearms and Tool Marks Exam iners (AFTE) conferences in 2005 and 2006.

Received 24 April 2008; and in revised form 5 Jan. 2009; accepted 11 Jan. 2009.

developed and implemented to generate the necessary tool mark signatures and well defined metrics were used to objectively evaluate the degree of similarity between known matching and nonmatching tool mark pairs. The distributions of the degree of similarity values obtained from the comparison of known matching and nonmatching pairs of tool marks were analyzed using established statistical techniques. While it is not possible to prove uniqueness statistically (14), the results of this study provide support for the concept that tool marks contain measurable features that exhibit a high degree of individuality.

Methods

The main goal of the study under consideration was to assess the degree of individuality and repeatability of the features transferred from the working surface of a tool to the tool marks created by it in an objective and repeatable manner. The approach selected to achieve this goal was by development of an automated tool mark comparison system. An automated comparison system provides both objective and repeatable results, since it applies the same algorithms and similarity metric to each tool mark pair under comparison. Moreover, such a system is capable of comparing large numbers of tool marks in a short period of time.

In addition to the development of an automated comparison system, a rigid methodology was formulated and followed for the creation of sample tool marks for the following three scenarios of interest:

Scenario (a) Comparison of tool marks when both the medium and the conditions under which different tool marks are created are the same.

Scenario (b) Comparison of tool marks when the conditions under which tool marks are created are the same, but the media are different.

Scenario (c) Comparison of tool marks when the medium onto which different tool marks are created is the same, but the conditions are different (this scenario was considered for the variations in the screwdriver's angle of attack only).

By analyzing the statistical distributions of similarity values resulting from the comparison of known matching and nonmatching pairs of tool marks, it is possible to assess the degree to which tool marks created by the same tool are repeatable and distinguishable from tool marks created by other tools. In this section, we provide an overview of the automated tool mark comparison system, the associated similarity metric, and the methodology followed for the creation of the tool mark samples used in this study.

3D Based Automated Tool Mark Comparison System

The implementation of an automated comparison system requires two main components: (i) data acquisition hardware and (ii) data analysis software. The data acquisition hardware is responsible for capturing the physical characteristics of the specimen being analyzed. The data analysis software is responsible for the storage, management, processing, and comparison of the data acquired by the data acquisition hardware. In the following sub sections, we describe these two components.

Data Acquisition Hardware From the inception of this study, it was decided that topographical images (often referred to as 3D data) as opposed to photographical images (referred to as 2D data) would be used to characterize the tool marks under comparison. Both topographical imaging and photographical imaging are processes which translate physical properties of the specimen into an array of numerical values. In the case of photographical images, these values correspond to the intensity of the light reflected by the specimen; in the case of topographical images, they correspond to the depth of the specimen's surface with respect to a reference plane. The use of topographical data has a number of important advantages over photographical data. Figure 1 shows an example of a topographical image on the left and a photographical image on the right corresponding to a striated tool mark created by a pair of tongue and groove pliers. Figure 1 demonstrates the vulnerability of photographical images to variations in the reflectivity of the medium onto which the tool mark is created. Other parameters which can influence photographical images are illumination conditions (intensity, angle, type of illumination, etc.), and camera angle. Topographical imaging is virtually immune to these variables, and therefore, provides a significantly more robust process to capture the relevant features of a specimen. In terms of flexibility, topographical data has the significant advantage of allowing for dimensionally preserving geometric transformations of the data. For example, topographical data can be mathematically "rotated" without distortion. This property plays an important role in the processing of the data. Figure 2 provides a visual representation of this characteristic. The images seen in Fig. 2 correspond to the same data, but from a different point of view. This is not always possible for photographical data (at least not accurately, unless multiple images are taken). Also, as seen in Fig. 2, topographical data allows for the identification and isolation of "waviness" (usually due to class characteristics) and "roughness" (often associated with individual characteristics). A more extensive discussion of the advantages of topographical data as opposed to photographical data can be found in (9).

There are a variety of technologies for the acquisition of topographical data that have been utilized in commercially available systems. For the purposes of this study, the candidate choices were constrained by the requirement that only noncontacting acquisition techniques be considered. The rationale for this requirement was that a contact based system would pose the risk of damaging the tool mark under consideration or altering the data if the same tool mark had to be acquired multiple times. At the start of this study, we considered several commercially available 3D imaging systems. These instruments utilize different technologies as indicated in Table 1. Among these, only the MicroSurf white light confocal microscope manufactured by NanoFocus AG (NanoFocus, Inc., Glen Allen, VA) and the NT series of white light interferometers manufactured by Veeco Instruments, Inc. (Chadds Ford, PA) provided the performance required for this project. Both these

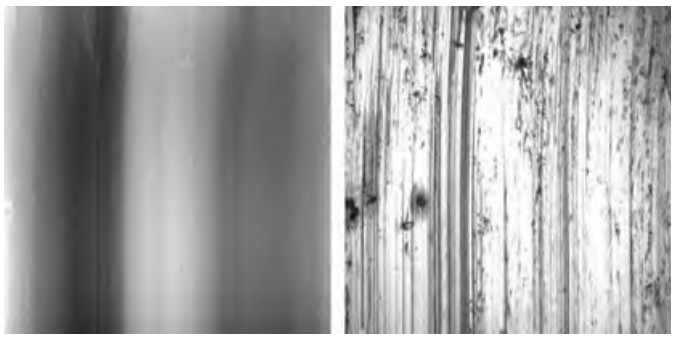

FIG. 1 *Example of topographical (left) and photographical (right) data for a striated tool mark.*

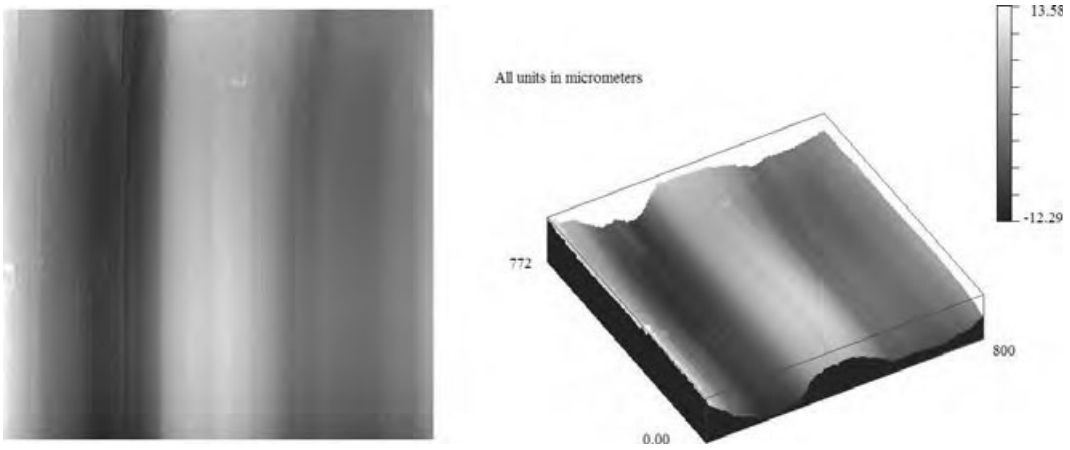

FIG. 2 *Geometric transformation of striated tool mark topographical data.*

TABLE 1 *Data acquisition systems evaluated.*

Manufacturer	Model	Technology	Data	Evaluation
LMI Technologies	LTS series	Triangulation	Single point	Inadequate lateral resolution
STIL	CHR	Chromatic Aberration	Single point	Inadequate parameters
NanoFocus	MicroScan	Dynamic Focusing	Single point	Limited range
Optimet	ConProbe/ConoLine	Conoscopic Holoraphy	Point/line	Inadequate lateral resolution
Veeco	NT series	WL Interferometry	Patch	Excellent performance
NanoFocus	MicroSurf	WL confocal microscope	Patch	Excellent performance

systems have exceptional lateral and depth resolution, and have the capability to acquire rectangular ''patches'' of points as opposed to single points or lines of data. The relative performance of the white light confocal microscope against the white light interferometer sensor is still a subject of debate within the metrology community. Nonetheless, there is evidence to suggest that the white light confo cal microscope can handle steeper slopes than its white light inter ferometer counterpart. On the other hand, the white light interferometer sensor may be able to achieve better depth resolution than the white light confocal microscope for relatively flat surfaces. Given that the lateral and depth resolution of the confocal micro scope was more than sufficient for the current application, that the slopes associated with tool mark topography are often significant, and that the cost of the confocal microscope was less than the white light interferometer, the NanoFocus MicroSurf white light confocal sensor was selected for our particular application. The operating conditions used in this study are shown in Table 2. The NanoFocus MicroSurf white light confocal microscope proved to be accurate, robust to vibration, and easy to use.

Data Analysis Software The automated comparison of data requires two main software components: the signature generation component, and the correlation component. The main purpose of the signature generation component is to isolate those features that are characteristic of the specimen under consideration (individual characteristics) from those that are common to all specimens of the same type (class characteristics). Consider, for example, the case of a group of screwdrivers of the same make and model. As these screwdrivers are manufactured to the same specifications, the over all geometric shape of the tool marks created by them is very simi lar. On the other hand, as no two manufactured parts are ever identical, there are microscopic variations specific to each screw driver blade. The key premise to be validated in this study is whether the process through which the blade features are trans ferred to a tool mark captures these specific features (most likely together with class characteristics features) in a repeatable manner. The challenge associated with the development of an effective automated tool mark comparison system is, therefore, to separate class characteristics from individual characteristics, and to treat them in the appropriate manner.

TABLE 2 *Main performance parameters of NanoFocus Microsurf.*

	Objective Lens	
	20×L	50×L
Numerical aperture	0.4	0.6
Single patch fov (μm)	800 × 800	320 × 320
Lateral resolution (μm)	1.5	0.6
Vertical resolution (nm)	20	10
Standoff (mm)	12	10.6

Signature Generation Component Figure 3 shows the main algorithmic modules of the signature generation component. These modules are:

Preprocessing: The unprocessed data obtained from the acquisi tion hardware is referred to as ''raw data.'' Raw data often includes inaccurate or questionable data points. We refer to such points as *unreliable* data points. The preprocessing module is responsible for the identification and preliminary handling of unreliable data points. Two types of unreliable data points are considered: drop offs and outliers.

Drop off points are points corresponding to regions of the speci men where the acquisition system has been unable to acquire data. In the case of optical systems, this limitation is generally because of insufficient light being collected by the optical system due to either low reflectivity or a steep slope on the specimen's surface. Such points are usually identified by the acquisition system as

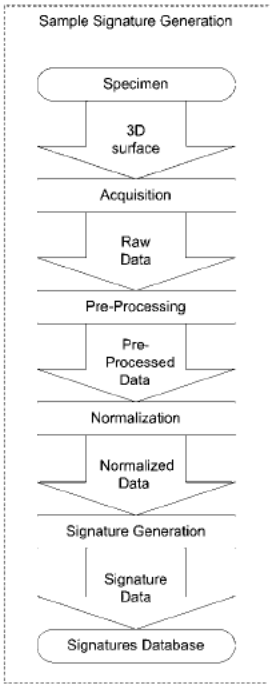

FIG. 3 *Signature generation steps.*

having insufficient light reflection intensity. As these points are identified by the optical system, there is no need to develop algorithms to recognize them. Nevertheless, the preprocessing software developed for this application keeps track of drop off points for later data handling.

"Outliers" are those data points that are inaccurately measured by the imaging system, but which are not recognized as such via the intensity of reflection information (in other words, the intensity of reflection associated with such points is within the nominal range). For this reason, these points are much more difficult to identify, and specific algorithms had to be developed for this purpose. Two approaches were used to identify such outliers. The first approach was based on the estimation of the slope between a point and its neighbors. Any point for which the local slope is above a preestablished threshold is identified as an outlier. The second approach was based on the statistical distribution of the data in the vicinity of the point under consideration. Any point which deviates beyond a predetermined number of standard deviations with respect to the local mean is considered an outlier. Once all unreliable points are identified, they are recorded in a "mask" which is then used for the remainder of the signature generation process. Figure 4 shows a raw topographical image of the same data as in Fig. 2 and its corresponding mask, where unreliable points are shown as dark points.

Normalization: The normalization module is responsible for compensating for the variations in the topographical images that result from inconsistencies during the acquisition process. A comprehensive presentation of the normalization process is beyond the scope of this paper. However, we consider a simple illustrative example. Let us assume that a given tool mark sample is acquired twice, but in each case, the tool mark surface is oriented differently. Figure 5 represents this situation, where a single cross section of data is considered for ease of presentation. Data 1 represents the data acquired the first time, while Data 2 represents the data acquired the second time. While these two sets of data correspond to the same tool mark (and should therefore be identical if one ignores instrument noise), they appear different due to a different relative orientation between the sample surface and the microscope during the acquisition process. If left uncorrected, these two data sets may be erroneously judged to be dissimilar by the correlation algorithms. The purpose of the normalization process is to bring these two data sets to a "level playing field." In the case of this example, the first step in the normalization process is to identify a baseline or a reference horizon. Let us assume that an appropriate baseline for the type of data under consideration is a linear function (in fact, the baseline could be a shape corresponding to a class characteristic). Once the baseline is identified, the purpose of the normalization is to apply a transformation to compensate for the fact that the tool marks under consideration were not acquired in a uniform manner. For the example under consideration, the simplest such transformation would be the rotation of the data.

Based on this simple example, we can articulate the purpose of the normalization process as it applies to any tool mark data of interest. The normalization process consists of the application of a geometric transformation to the preprocessed data in an effort to compensate for any inconsistencies resulting from the acquisition process. In other words, the goal of the normalization process is to ensure that the data is represented in a consistent way regardless of variations which may have taken place during the acquisition process.

It is important to note that the normalization process would be significantly more challenging if not impossible if the data under consideration were photographical data. While processes similar to normalization can be developed for photographical data, it would be significantly more difficult to achieve the same level of accuracy as that achievable with topographical data. Also, it is worth noting that in order to perform the normalization process accurately, it is necessary to have knowledge of which points can be considered reliable. For the example under consideration, only reliable points are used in the estimation of the baseline. Otherwise,

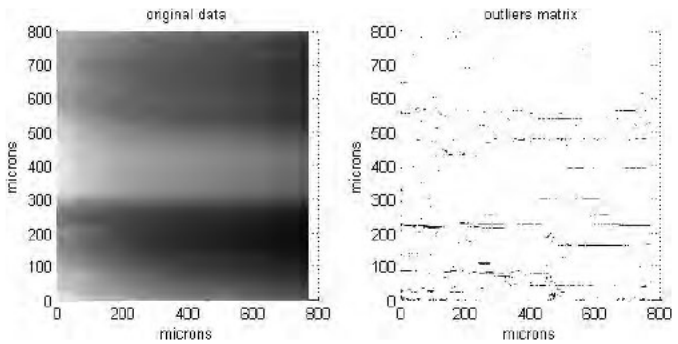

FIG. 4 *Tongue and groove pliers tool mark (left) and corresponding mask (right). Unreliable points are indicated as dark in the mask.*

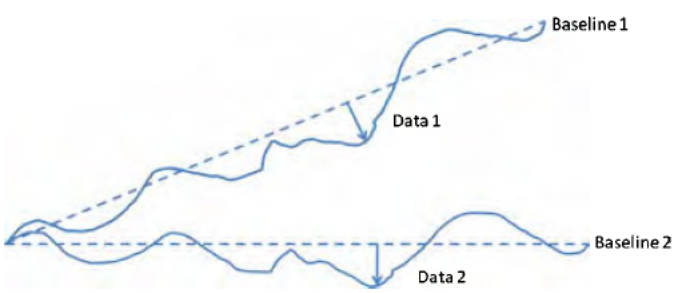

FIG. 5 *Conceptual example of normalization process in the case of different orientations.*

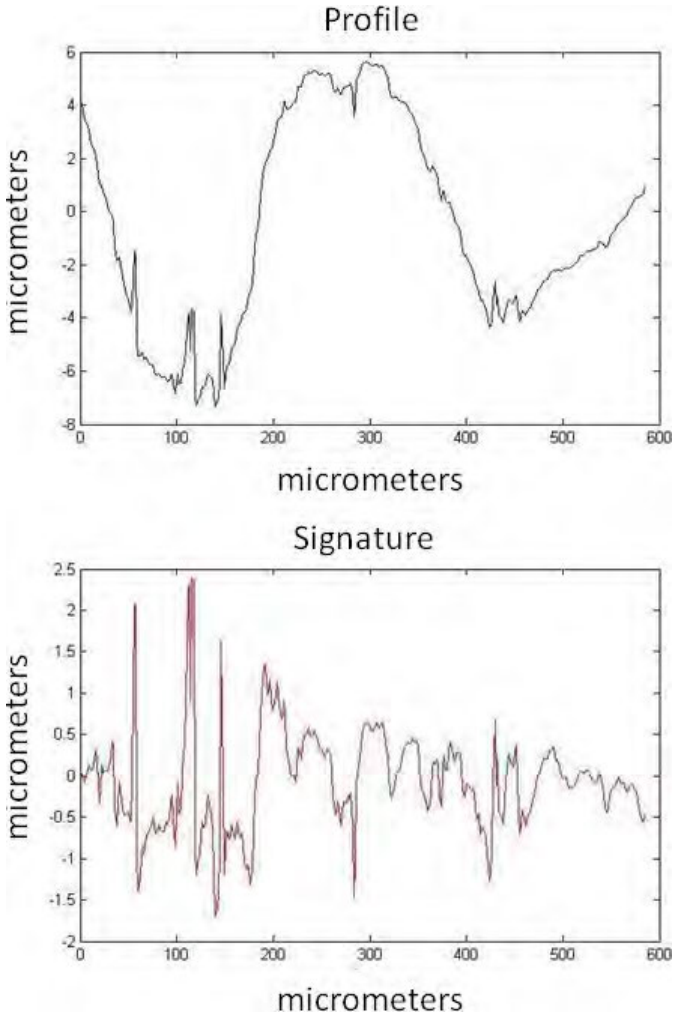

FIG. 6 *Profile (top) and signature (bottom) of the tool mark.*

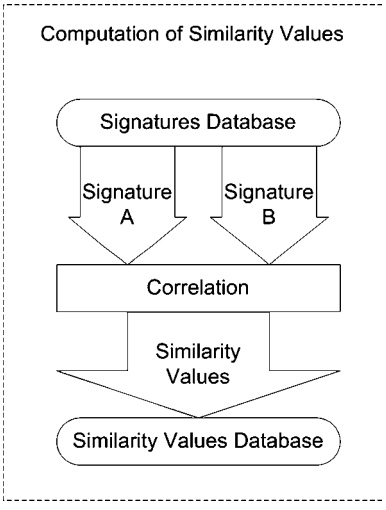

FIG. 7 *Signature correlation steps.*

the result of the normalization process is not consistent between different tool marks. For this reason, identification of unreliable points precedes the normalization process.

Signature Generation: The signature generation module is responsible for emphasizing those features which are specific to the tool mark under consideration (individual characteristics), while minimizing the features which may be common to all tool marks of the same type (class characteristics). For the tool marks under consideration, this process consists of two steps. The first step involves the conversion of the topographical tool mark data (in the form of a 2D array) into a single data vector that corresponds to a cross section of the tool mark. The second step involves using a Gaussian band pass filter to eliminate the low frequency component corresponding to the class characteristics of the tool mark. Figure 6 shows an example of the signature generation process applied to normalized data profile. Notice that as a result of the signature gen eration process, all low frequency components are discarded, while the high frequency components are left intact.

Signature Correlation Component The signatures generated by the signature generation module are stored in a database, and are accessible to the signature correlation component (see Fig. 7). Given a pair of signatures, the purpose of the signature correlation component is to evaluate a metric indicative of their degree of sim ilarity. We refer to the value achieved by such metric as a *similar ity measure*. Let us denote a pair of signatures corresponding to two different striated tools mark by:

$$z_i(n), z_j(n); n = 1, ..., N. \tag{1}$$

where the mean value of z_k (denoted by z_k) is equal to zero for both $k = i$, j. We define the *relative distance* between two signatures of the same number of points as:

$$\mathrm{rdist}_{i,j} = 1 - \frac{\sum_{n\ 1,...,N} (z_i(n) - z_j(n))^2}{\sum_{n\ 1,...,N} (z_i(n) + z_j(n))^2} \tag{2}$$

The relative distance metric is a time domain similarity metric (as opposed to frequency domain, wavelet domain, etc.), and it offers advantages in terms of being well suited to handle signatures of different lengths and signatures with missing data points (unreli able data points). The relative distance defined in (2) is upper bounded by 1, where a relative distance of 1 indicates that the two signatures satisfy $z_i(n) = z_j(n)\ \forall n = 1, ..., N$. In other words, the two signatures are identical. On the other hand, a similarity metric value close to zero indicates that there is only a subtle (or insignifi cant) relationship between the two signatures.

As discussed with reference to the normalization process, it is reasonable to assume that there will be differences in the area imaged for each tool mark. For this reason, while computing the similarity measure between two signatures, it is necessary to allow a pre established degree of relative lateral displacement or "shift" between them. However, as one signature is "shifted" with respect to its counterpart, the number of points of comparison decreases. For this reason, a slight modification of Equation (2) is necessary. Let us consider the case where signature j is shifted to the right by Δ points with respect to signature i. In such a case, the number of overlapping points between the two signatures decreases to $N - \Delta$, and the region of overlap between the two signatures becomes:

$$z_i(n \quad \Delta), z_j(n); n = 1, ..., N - \Delta \tag{3}$$

The relative distance between the two shifted signatures is com puted by:

$$\mathrm{rdist}_{i,j}(\Delta) = 1 - \frac{\sum_{n\ 1,\ldots,N\ \Delta} (z_i(n\quad \Delta) - z_j(n))^2}{\sum_{n\ 1,\ldots,N\ \Delta} (z_i(n\quad \Delta) + z_j(n))^2} \qquad (4)$$

A similar computation can be made in the case of a left shift, which is denoted by a negative value of Δ. Based on this definition, the similarity measure between two signatures is defined by:

$$s_{i,j}(\Delta_{\max}) = \max_{|\Delta|<\Delta_{\max}} \mathrm{rdist}_{i,j}(\Delta) \qquad (5)$$

The maximum relative shift $\Delta_{\max}$ in Equation (5) is selected so as to reflect the inconsistencies inherent to the acquisition process. The properties of the similarity metric defined by Equation (5) are inherited from the properties of Equation (2).

Tool Selection and Sample Tool Marks Preparation

While both striated and impressed tool marks were considered as part of this study, this paper only discusses striated tool marks (the results obtained for impressed tool marks are in preparation). In particular, we consider two types of tools: screwdrivers and tongue and groove pliers. The screwdrivers used in this study were Craftsmen Professional screwdrivers (model # 47441) while the tongue and groove pliers used in this study were Cooper Tools Crescent pliers (model # R210C).

Tool Marks Sample Preparation For each of the tool types under consideration, 10 sample tools of the same manufacturer and model number were purchased. For each sample tool, 10 tool mark samples were created under the same conditions for each medium of interest. We refer to each such group of 100 tool marks created on the same medium and under the same conditions as a *set*. Table 3 summarizes the different sets of tool mark samples created as part of this study. As shown in Table 3, seven different sets of tool marks, totaling 700 individual specimens were used for this study.

While creating the sample tool marks, care was taken to minimize the likelihood of damaging the working surface of the tool. For this reason, the first set of tool mark samples was created on lead for both tool types. Once the repeatability and individuality of these tool marks was evaluated, we proceeded to harder media. In the case of screwdrivers, sample tool marks at three angles of attack (15 , 30 , and 45) were created in lead, and an additional set was created at an angle of attack of 30 on aluminum. The cross sectional width of these striated tool marks was 5.0 mm. In the case of the tongue and groove pliers, the creation of tool mark samples in lead rope was followed by the creation of samples on brass and galvanized steel pipe. The cross sectional width of these tool marks was *c.* 7.4 mm.

TABLE 3 *Tool mark sets.*

Set	Tool Type	Conditions	Media
SD01	Screwdriver	45 deg	Lead
SD02	Screwdriver	30 deg	Lead
SD03	Screwdriver	15 deg	Lead
SD04	Screwdriver	30 deg	Aluminum
TG01	Tongue and groove pliers		Brass
TG02	Tongue and groove pliers		Galvanized steel
TG03	Tongue and groove pliers		Lead

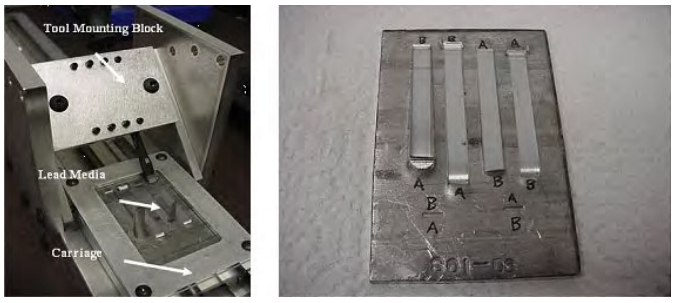

FIG. 8 *Device to create screwdriver striated tool marks (left) and an example of such a tool mark (right).*

Creation of Tool Marks from Screwdrivers In the case of screwdrivers, we found that it was very difficult to create uniform striations manually. For this reason, a device to assist in the creation of these tool marks was designed and built (see left side of Fig. 8). The main components of this device are a carriage and a tool mounting block. The carriage was designed so that a 5.08 × 7.62 cm piece of metal sheet could be rigidly affixed to it. The tool mounting block was designed such that a screwdriver could be rigidly mounted at a variety of predetermined angles with respect to the medium affixed to the carriage. The carriage could then be translated using a lead screw, allowing for the displacement of the medium with respect to the blade of the screwdriver. Furthermore, the lead screw was motorized using a conventional electric drill, producing a constant speed displacement of the sample medium with respect to the screwdriver blade. This device enabled the creation of very clean and uniform tool marks. With the assistance of this device, the sample tool mark sets were created on both lead and aluminum sheets in an identical fashion. An example of the types of sample tool marks created with this device can be seen on the right side of Fig. 8.

Prior to making the test samples, each screwdriver was labeled with an identifying number between 01 and 10. Also, both sides of the screwdriver blade were labeled using the conventional A B labeling (see Fig. 9). The screwdriver tool mark samples were created on lead and aluminum. 30.48 × 30.48 × 0.32 cm sheets were cut into 5.08 × 7.62 × 0.32 cm rectangles using a metal shear. Prior to labeling the medium, each sample was flattened by impacting it with a dead blow hammer. The medium was then labeled "SXX YY" along the bottom edge, where "XX" refers to the screwdriver's label (01 through 10) and "YY" refers to the tool mark sample number (01 through 40). The sample was placed in the aluminum frame and held in place by securing the upper frame plate with four screws. It was then positioned so that the screwdriver, when fixtured, contacted the upper region of the sample. The screwdriver was placed into the tool mounting block with the

FIG. 9 *Steps involved in the creation of tool marks from screwdrivers.*

FIG. 10 *Steps involved in the creation of tool marks from tongue and groove pliers.*

blade labeled "A" facing upwards. The end of the screwdriver was then gently tapped until the tip just impacted the surface of the medium sample. It was then secured with two set screws. By operating the drill, the sample was moved toward the rear of the device for about an inch, creating a "push" tool mark on the sheet. Push and pull marks were made with each side of the blade and two impression marks were made using the tip of the screwdriver as seen in the right side of Fig. 8. For this study, only one push mark from each tool impression sample was used. After the creation of the tool mark, the samples were stored in a container. Special care was taken at every step of the process to avoid contact with skin or moisture so as to minimize the oxidation rate.

To study the effect of variation of screwdriver angle of attack on striated tool marks, a total of 300 striated tool marks were created on lead at screwdriver angle of attack of 15 , 30 , and 45 to the lead sheet (see Table 3). To study the effect of media, 100 tool mark samples were created at a screwdriver angle of attack of 30 on aluminum sheet using the same procedure. Topographical images of all of the prepared tool marks were acquired with a lateral resolution of 1.52 μm. For processing purposes, these data sets were decimated to a lateral resolution of 4.56 μm.

Creation of Tool Marks from Tongue and Groove Pliers For the creation of striated tool marks from tongue and groove pliers, it was decided that the tool mark of interest would correspond to the striated tool mark created by a single predetermined tooth on the jaw of the tongue and groove pliers. Each tool mark was created by the rotation of the tool as the jaws firmly grip the curved surface of a cylindrical sample of the medium in a uniform fashion. A description of the steps involved in the creation of three sets of 100 tool marks on brass pipes, steel pipes, and lead rope follows (see Fig. 10).

Prior to making the tool marks, each pair of tongue and groove pliers was labeled with an identifying number between 01 and 10. Both sides of both jaws of each pair of the tongue and groove pliers were labeled using the traditional a b and A B labeling convention. The tooth of interest on the jaw of the tongue and groove pliers, which on contact with the pipe/rope would create the striated tool mark, was identified and marked. The indication of the tooth of interest was made with a punch. Appropriate care was taken to ensure that this process did not physically alter the tooth in any way. The media to be used for the creation of the striated tool marks from tongue and groove pliers were brass pipes, galvanized steel pipes, and lead rope. The pipes/ropes were selected to have a 1.27 cm internal diameter, which facilitated the contact of the same tooth of the tool's jaw for all media of interest. The lead rope was cut into equal pieces of 5.08 cm, while the brass and galvanized steel pipes were cut into equal pieces of 25.40 cm using a band saw. The pipe/rope was rigidly mounted on to a vise affixed to a work bench and was tightly clamped to eliminate movement during the creation of the tool mark. The jaw of the tongue and groove pliers was brought in contact with the pipe/rope to identify the region on the pipe/rope where the tooth of interest would make contact. Once the tooth was aligned satisfactorily over the surface of the medium, the region of contact on the pipe/rope was identified by marking it with a line drawn with a soft tip marker. The purpose of drawing this line was to indicate the position on the pipe/rope where the tooth of interest would be initially placed to create the tool mark. The tongue and groove pliers were brought in contact with the surface of the pipe/rope so that the tooth of interest was in alignment with the line drawn in the previous step. While holding the tongue and groove pliers firmly with both hands, it was slowly rotated around the surface of the pipe/rope such that only the tooth of interest was in direct contact with the pipe/rope over the region of interest. This rotational movement resulted in the creation of a tool mark consisting of striations imparted from the movement of the tooth of interest over the pipe's/rope's surface. Once a sufficiently long striated tool mark was created (about half to 1 cm in length), the tongue and groove pliers were carefully withdrawn from the pipe's/rope's surface. After the tool mark had been created, a soft brush was used to clean the jaws of the tongue and groove pliers before the creation of the subsequent tool mark. Each tool mark was labeled as "TSXX YY" where "XX" referred to the tongue and groove pliers (01 through 10) and "YY" referred to the tool mark sample number (01 through 30). The brass and galvanized steel pipes were then cut into two pieces by a band saw such that each half had five tool marks and then stored in a container.

Three dimensional images of each of the prepared tool marks were acquired with a lateral resolution of 1.52 μm. For processing purposes, these data sets were decimated to a lateral resolution of 4.56 μm.

Statistics

Tables 4 and 5 summarize the sets of data which were compared, and the number of matching (i.e., same tool) and nonmatching (i.e., different tool) comparisons performed for screwdrivers and tongue and groove pliers tool marks, respectively. Each set of comparisons shown in Tables 4 and 5 corresponds to one of the three scenarios discussed in the Methods section. As an example,

TABLE 4 *Numbers of comparisons of matching/nonmatching pairs of screwdriver tool marks.*

Matching/ Nonmatching Pairs	SD01	SD02	SD03	SD04
SD01	450/4500	1000/9000	1000/9000	X
SD02		450/4500	1000/9000	1000/9000
SD03			450/4500	X
SD04				450/4500

TABLE 5 *Numbers of comparisons of matching/nonmatching pairs of tongue and groove pliers tool marks.*

Matching/ Nonmatching Pairs	TG01	TG02	TG03
TG01	450/4500	1000/9000	1000/9000
TG02		450/4500	1000/9000
TG03			450/4500

consider the comparison of set SD01 against itself. Such comparison resulted in 450 matching similarity measure values and 4500 nonmatching similarity measure values of tool marks created onto the same medium, under the same conditions. The comparison of set SD01 against itself corresponds to Scenario (a). Set SD01 was also compared against set SD02, resulting in 1000 matching similarity measure values and 9000 nonmatching similarity measure values of tool marks created onto the same medium, under different angle of attack. This set of comparison corresponds to Scenario (c). By analyzing the differences between the distributions obtained from the comparisons of SD01 versus SD01, SD02 versus SD02, and SD01 versus SD02 it is possible to isolate and evaluate the effect of screwdriver angle of attack on the created tool mark. In a similar manner, the effect of different media was analyzed (Scenario [b]).

The purpose of performing the large number of correlations discussed above is to empirically estimate the distribution of matching and nonmatching similarity measure values for the scenarios of interest. An analysis of these distributions allows us to conclude whether the tool marks created by these tools under the conditions of interest display individualizing and repeatable features. If these distributions are distinct at a given level of significance, we can conclude that the individuality and repeatability criteria have been verified, or at least have not been disproven to that level of significance. Figure 11 shows the empirically estimated matching and nonmatching similarity measure distributions for screwdriver tool marks created on lead at a 30° angle of attack (set SD02). As seen in Fig. 11, the distributions of matching and nonmatching similarity measure values are quite distinct. The nonmatching distribution has a mean of .33 with a standard deviation of .07, while the matching distribution has a mean of .92 with a standard deviation of .07. Clearly, these empirical distributions indicate a high degree of similarity among marks from the same tool (repeatability) and differences between marks from different tools (individuality). The same behavior can be observed in the inter comparison of sets SD02, SD03, and SD04 corresponding to the comparison of screwdriver tool marks, and TG01, TG02, and TG03 corresponding to the comparison of tongue and groove pliers tool marks. In all these cases, either no or minimal overlap can be seen between the distributions. As an example of these results in the case of tongue and groove pliers, Fig. 12 shows the matching and nonmatching similarity distributions for tongue and groove tool marks created on steel pipes (set TG02).

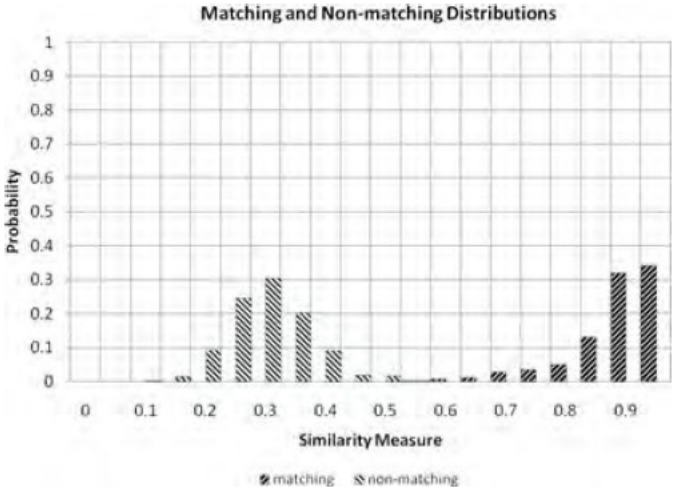

FIG. 11 *Matching and nonmatching distributions of similarity values for screwdriver striations on lead sheet at 30 degrees.*

To summarize the behavior of each of the sets of comparisons shown in Tables 3 and 4, it is convenient to select a metric which quantifies the degree of overlap between the matching and nonmatching distributions. Such a simple and convenient metric is the *empirical error rate*. The empirical error rate is a simple metric which has the appealing feature of having an intuitive interpretation. A brief description of this metric follows:

Empirical Error Rate

Having the empirically generated distributions of matching and nonmatching similarity values, it is possible to compute an optimal threshold such that if a given pair of tool marks yields a similarity value above such threshold, it is assumed that the pair of tool marks under comparison match. Similarly, if a given pair of tool marks yields a similarity measure below the optimal threshold, it is assumed that the pair of tools marks under comparison does not match. The boundary or threshold value is selected to minimize the empirical error rate (defined as the mean of both false positive and

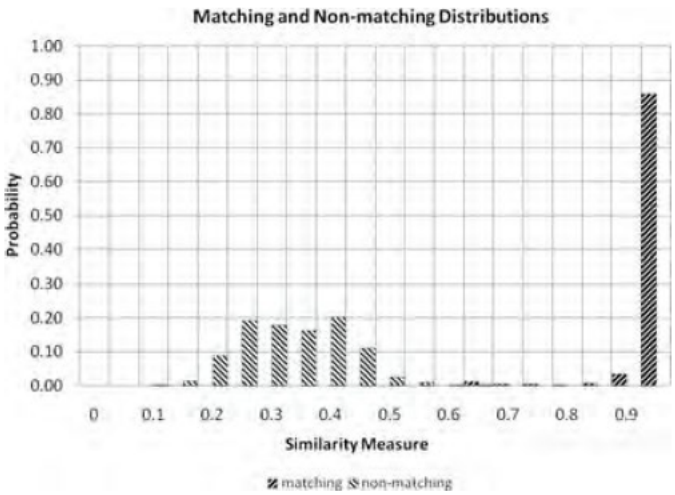

FIG. 12 *Matching and nonmatching distributions of similarity values for tongue and groove pliers striations on steel pipes.*

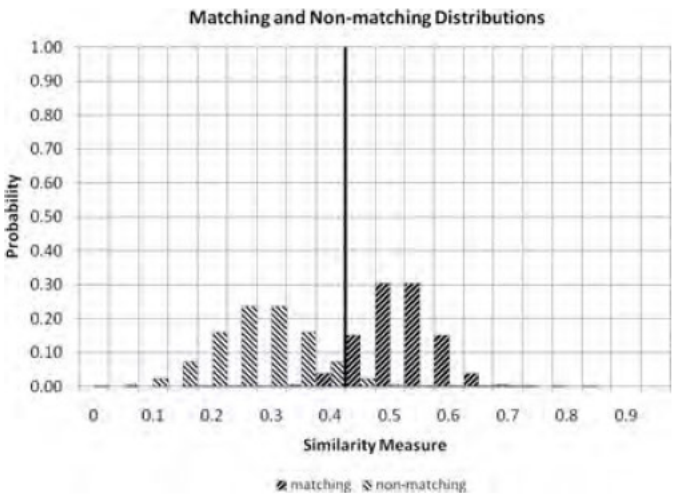

FIG. 13 *Empirical error rate estimation.*

false negative probabilities of error). We could have selected any other threshold value so as to shift the proportions of each type of error as desired. Figure 13 shows a graphical representation of this approach, where two distributions are shown a matching distribu tion and a nonmatching distribution. Having identified the optimal threshold (vertical line), it is possible to estimate the probability of false positive and false negative identification. We use the empirical error rate as a metric of tool mark individuality, where a low empirical error rate is indicative of high specificity and repeatability.

It is important to note that the empirical error rates obtained as part of this study depend not only on the repeatability and individu ality of the tool marks under consideration but also on the algo rithms developed as part of the automated comparison system. These algorithms are significantly less sophisticated than the pattern recognition capabilities of a well trained human tool mark exam iner. Therefore, while the results presented in this paper have the benefit of objectivity, they are not meant to provide an estimate of the probability of an erroneous identification by an experienced tool mark examiner.

Results and Conclusions

In this section, we present the results obtained in each of the three scenarios described in the Methods section.

Scenario (a) Same Medium, Same Conditions

The empirical error rates for all screwdriver tool mark compari sons are summarized in Table 6. Among these results, the ones that correspond to Scenario (a) are located along the diagonal of the table (i.e., comparisons of SD01 vs. SD01, SD02 vs. SD02, SD03 vs. SD03, and SD04 vs. SD04). In all but one case, the empirical error rate is 0.00%. The only exception corresponds to SD01 versus SD01, where the empirical error rate 0.11% corresponds to a false exclusion out of 450 matching comparisons and no false inclusions out of 4500 nonmatching comparisons. There are no incorrect matches of two different tools for any of the same medium, same angle comparisons. These results indicate that for the media and angles of attack under consideration, the resulting screwdriver tool marks are sufficiently repeatable and specific to allow for very reli able identification. It may be significant that the only observed errors are at the highest angle of attack.

In a similar manner, Table 7 summarizes the results for tongue and groove pliers. As in the case of screwdrivers, those which cor respond to Scenario (a) are located along the diagonal of the table (i.e., comparisons of TG01 vs. TG01, TG02 vs. TG02, and TG03 vs. TG03). Once again, in all but one case, the empirical error rate is 0.00%. The only exception corresponds to TG01 versus TG01, where the empirical error rate 0.03% corresponds to no false exclusions out of 450 matching comparisons, and three false inclusions out of a total of 4500 nonmatching comparisons. As for the screwdriver marks, these results indicate that for the media under consideration, the tongue and groove pliers tool marks are sufficiently repeatable and specific to allow for very reliable identification. The effect of the metals studied does not appear to be significant, since all metals produce very low error rates and the only errors observed are on brass which has hardness intermediate between that of lead and steel.

TABLE 6 *Empirical error rate for screwdriver tool mark comparisons.*

			Lead			Aluminum
			45deg	30deg	15deg	30deg
Empirical Error Rate			SD01	SD02	SD03	SD04
Lead	45deg	SD01	0.11%	13.61%	49.50%	X
	30deg	SD02		0.00%	33.51%	8.36%
	15deg	SD03			0.00%	X
Aluminum	30deg	SD04				0.00%

TABLE 7 *Empirical error rate for tongue and groove pliers tool mark comparisons.*

		Brass	Steel	Lead
Empirical Error Rate		TG01	TG02	TG03
Brass	TG01	0.03%	0.23%	2.46%
Steel	TG02		0.00%	1.58%
Lead	TG03			0.00%

Scenario (b) Different Media, Same Conditions

Table 6 includes the empirical error rates resulting from the comparison of screwdriver tool marks created under the same con ditions (30 of attack) but onto different media (lead vs. aluminum: sets SD02 vs. SD04). As discussed for Scenario (a), the empirical error rate is 0.00% when screwdriver tool marks created onto the same medium at an attack angle of 30 are compared for both alu minum and lead. As shown in Table 6, it increases to 8.36% when tool marks on lead are compared with tool marks on aluminum. This 8.36% error rate corresponds to 63 false exclusions out of 1000 matching comparisons and 938 false inclusions out of 9000 nonmatching comparisons.

In a similar fashion, Table 7 includes the empirical error rate resulting from the comparison of tongue and groove pliers tool mark samples created in different media (comparisons TG01 vs. TG02, TG01 vs. TG03, and TG02 vs. TG03). For striation marks produced by tongue and groove pliers the medium onto which the tool marks are created has a measurable effect on the tool marks. The empirical error rate for brass versus steel comparison is rela tively low at 0.23%, corresponding to represent four false exclu sions out of 1000 matching comparisons and six false inclusions out of 9000 nonmatching comparisons. The reasonable success rate for these two metals probably results from the fact that they do not differ greatly in hardness. In contrast, comparison of marks on either brass or steel with those on lead result in higher error rates, 2.46% and 1.58%, respectively.

Scenario (c) Same Medium, Different Conditions (Screwdrivers Only)

Table 6 also includes the empirical error rate resulting from the comparison of screwdriver tool marks created on the same medium (lead) but under different angles of attack. As shown in Table 6, the variation of the angle of attack has a significant effect on the resulting tool mark even if the medium is the same. The error rates for comparison increase as the difference between the angle of attack is increased. The total error rates are pronounced enough that comparison of tool marks created at 15 with those created at 45 is no better than random guessing, which would have an error rate of 50% (close to the obtained 49.5%). The likely reason for the inability to correctly match tool marks made by the same tool at different angles of attack is that the points of the tool surface that

are in contact with the receiving surface are different at the two angles.

Discussion

As stated at the beginning of this paper, the main goal of the study herein reported is to validate the basic premise of tool mark identification. As can be seen, the results obtained from this study provide substantial evidence to the validity of this basic premise of tool mark identification in the case of striated tool marks.

A number of important conclusions can be derived from the results, discussed in the previous section, as stated below:

- Striated tool marks produced by screwdrivers and tongue and groove pliers are both repeatable and specific enough to allow for reliable identification of the producing tool when they are created on the same medium and under the same conditions (for the media and tools used in this evaluation).
- When striated tool marks are created on different media but under the same conditions, the tool marks can still be identified with high reliability. In the case of tongue and groove pliers, it is interesting to note that the empirical error rate increases with an increase in the degree of dissimilarity in the hardness of the medium onto which the tool marks are created. This implies that while the practice of creating control tool marks in lead is a sound one from the perspective of avoiding damage to the tool's working surface, a higher degree of agreement may be achievable if tool marks are created onto media of similar hard ness as that of the evidence tool mark.
- Screwdriver striated tool marks depend significantly on the angle of attack at which the tool mark is created (more so than with respect to the media). So much so, that tool marks created by the same screwdriver may appear completely different if cre ated at drastically different angles of attack. Therefore, the com parison of an evidence screwdriver tool mark requires the creation of control tool marks at multiple angles of attack.
- It was observed that irrespective of the type of comparison (i.e., within the same sets such as TG01 vs. TG01 or between differ ent sets such as TG01 vs. TG02, etc.), the nonmatching distribu tions obtained for a given tool type always had similar characteristics, in particular a low median and relatively low standard deviation. While this is not a surprising result, it has meaningful implications. First of all, it provides strong evidence to the premise that the probability of obtaining a high degree of similarity while comparing a pair of nonmatching tool marks is extremely low. If the behavior observed for the set of tools used in this evaluation can be considered as characteristic of all tools of the same type (which is likely to be the case at least for those tools manufactured by the same techniques), the probabil ity of a pair of different tools having similar features is extre mely low.
- It was observed that in some of the cases where both the condi tions and media were the same (Scenario a) the empirical prob ability of error was not always zero. Upon inspection of the raw tool mark images, it was noticed that the nonzero probability of error was because of the presence of a very small number of "bad" tool mark images (where we loosely use the term bad to indicate that such tool marks display highly anomalous features as a result of the creation and/or acquisition process). These bad images resulted in a matching pair being erroneously classi fied as a nonmatching pair (and never the other way around). In other words, the probability of error originated from a faulty image, and not because the tool itself would not create repeat able and individual tool marks (as other tool marks created by the same tool resulted in perfectly good images). Given the low probabilities of error associated with these cases, even a single bad tool mark image can have a relatively significant effect.

Based on these observations, it is evident that the obtained results provide substantial evidence to the validity of the basic premise of tool mark identification. Furthermore, these results reinforce the validity of many current practices of tool mark examiners.

Acknowledgments

Funded by the Federal Bureau of Investigation under Contract Number J FBI 02 128. Mention of trade names is for information purposes only and does not imply endorsement by the FBI or the federal government.

References

1. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
2. Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).
3. Davis JE. An introduction to tool marks, firearms and the striagraph. Springfield, IL: Charles C Thomas Pub Ltd, 1958.
4. De Kinder J, Prevot P, Pirlot M, Nys B. Surface topology of bullet stria tions: an innovating technique. AFTE J, Spring 1998;30(2):294 9.
5. De Kinder J, Bonifanti M. Automated comparison of bullet striations based on 3D topography. Forensic Sci Int 1999;101(2):85 93.
6. Geradts ZJ, Zaal D, Hardy H, Lelieveld J, Keereweer I, Bijhold J. Pilot investigation of automatic comparison of striation marks with structured light. Proc SPIE 2001;4232:49 56.
7. Bachrach B. Development of a 3D based automated firearms evidence comparison system. J Forensic Sci 2002;47(6):1253 64.
8. Banno A, Masuda T, Ikeuchi K. Three dimensional visualization and comparison of impressions on fired bullets. Forensic Sci Int 2004;140(3):233 40.
9. Miller J. An introduction to the forensic examination of toolmarks. AFTE J, Summer 2001;33(3):233 48.
10. Biasotti A, Murdock J. Criteria for identification or state of the art of firearm and toolmark identification. AFTE J 1984;16(4):16 34.
11. Miller J, McLean M. Criteria for identification of toolmarks. AFTE J 1998;30(1):15 61.
12. Nichols RG. Firearm and toolmark identification criteria: a review of the literature. J Forensic Sci 1997;42(3):466 74.
13. Nichols RG. Firearm and toolmark identification criteria: a review of the literature, part II. J Forensic Sci 2003;48(2):318 27.
14. Stoney DA. What made us ever think we could individualize using sta tistics? J Forensic Sci 1991;31(2):197 9.

Additional information and reprint requests:
Benjamin Bachrach, Ph.D.
Vice President
Director, Electro mechanical Systems
Intelligent Automation, Inc.
15400 Calhoun Drive, Suite 400
Rockville, MD 20855
E mail: bach@i a i.com

JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, July 2010, Vol. 55, No. 4
doi: 10.1111/j.1556 4029.2010.01424.x
Available online at: interscience.wiley.com

PAPER

CRIMINALISTICS

L. Scott Chumbley,[1] Ph.D.; Max D. Morris,[1] Ph.D.; M. James Kreiser,[2] B.S.; Charles Fisher,[1] B.S.; Jeremy Craft,[1] M.S.; Lawrence J. Genalo,[1] Ph.D.; Stephen Davis,[1] B.S.; David Faden,[1] B.S.; and Julie Kidd,[1] M.S.

Validation of Tool Mark Comparisons Obtained Using a Quantitative, Comparative, Statistical Algorithm

ABSTRACT: A statistical analysis and computational algorithm for comparing pairs of tool marks via profilometry data is described. Empirical validation of the method is established through experiments based on tool marks made at selected fixed angles from 50 sequentially manufactured screwdriver tips. Results obtained from three different comparison scenarios are presented and are in agreement with experiential knowledge pos sessed by practicing examiners. Further comparisons between scores produced by the algorithm and visual assessments of the same tool mark pairs by professional tool mark examiners in a blind study in general show good agreement between the algorithm and human experts. In specific instances where the algorithm had difficulty in assessing a particular comparison pair, results obtained during the collaborative study with professional examin ers suggest ways in which algorithm performance may be improved. It is concluded that the addition of contextual information when inputting data into the algorithm should result in better performance.

KEYWORDS: forensic science, tool mark comparison, comparison microscope, screwdriver, statistics, striae

In the fifteen years since the 1993 *Daubert versus State of Florida* decision, increasing attacks have been aimed at firearm and tool mark examiners by defense attorneys via motions to exclude evidence based on expert testimony. Such motions claim that the study of tool marks has no scientific basis, that error rates are unknown and incalculable, and that comparisons are subjective and prejudicial. Often persuasive, these motions skillfully blend truth with unsupported assertions or assumptions in a number of ways. First, the claim that scientific evidence is lacking in tool mark examinations ignores the numerous studies that have been con ducted, especially in the area of firearms (1 4), to investigate the reproducibility and durability of markings. These studies have shown time and again that while matching of cartridges cannot be universally applied to all makes and models of guns using all types of ammunition, the characteristic markings produced are often quite durable and a high percentage can be successfully identified using optical microscopy. Second, the claims that error rates are unknown, and that the probability of different guns having identical markings has not been established, are true. However, it must be understood that establishing error rates and probabilities in the area of tool marks is fundamentally different than in an area such as genetic matching involving DNA. When considering genetic match ing, all the variables and parameters of a DNA strand are known and error rates can be calculated with a high degree of accuracy. This is not the case in tool marks where the variables of force, angle of attack, motion of the tool, surface finish of the tool, past history of use, etc. are not known or cannot be determined, and the possibility for variation is always increasing as the population under study continues to increase and change. For practical purposes, this may indeed mean that realistic error rates cannot be completely characterized, but experiments based on sequentially manufactured tools may lead to useful approximations and/or bounds.

Finally, it is also true that an examiner necessarily offers a sub jective opinion when rendering a decision. However, the pattern on which that decision is based consists of striations that can be char acterized and quantified in an objective, mathematical manner. The proposition that tool marks must necessarily have a quantifiable basis is the principle upon which the Integrated Ballistics Imaging System (IBIS) developed and manufactured by Forensic Technol ogy, Inc. for bullets and cartridge cases operates. IBIS uses fixed lighting and an image capture system to obtain a standard digital image file of the bullet or cartridge case. The contrast displayed in the image is reduced to a digital signal that can then be used for rapid comparisons to other files in a search mode. The latest ver sion of IBIS uses the actual surface roughness as measured by a confocal microscope to generate a comparison file. The results are displayed in a manner analogous to a web search engine, where possibilities are listed in order with numbers associated with each possibility. An experienced tool mark examiner must then review the list of possibilities to make a judgment as to whether a match does, in fact, exist. In instances where a match is declared, it is quite common for the match not to be the first possibility displayed by IBIS, but to be further down the list. In other words, while the analysis/algorithm employed by FTI produces the numbers associ ated with each match, these numbers carry no clear statistical

[1]Ames Laboratory, Iowa State University, 2220 Hoover, Ames, IA 50011.
[2]Illinois State Police, Retired, 3112 Sequoia Dr., Springfield, IL 62712.

Received 5 Feb. 2009; and in revised form 12 April 2009; accepted 19 April 2009.

relevance or interpretation related to the quality or probability of match of any given comparison (5). However, as the marks under investigation can be quantified, there appears to be a significant potential for advancement in analyses of such data. An objective method of analysis should be possible for any given type of tool mark, and (at least in principle) an error rate established for com parisons made between any given subset of marks within a larger population of similar marks.

Researchers at Iowa State University have developed a com puter based data analysis technique that allows rapid comparison of large numbers of data files of the type that might be produced when studying striated tool marks A major aim of the research reported here is to construct well defined numerical indices, based upon the information contained within the tool mark itself, that are useful in establishing error rates for objective tool mark matching. While this error rate may only be practically achievable for a par ticular set of experimental conditions, it should serve as a bench mark error rate for subsequent studies. Initial results (6) indicated that simple statistics computed from the quantitative data produced by a surface profilometer, namely, maximized data correlations over short data segments, supported the empirical assertions of forensic examiners concerning comparisons of tool marks generated on lead plates by consecutively manufactured screwdriver tips. One drawback in using maximized correlations is that there is no clear standard against which they can be objectively compared. In some cases, maximized correlations may be high, implying a high degree of linear agreement between data pairs, but not necessarily imply ing strong similarity between the tool mark patterns. In others, the linear correlations over short data segments may be smaller, but the overall tool mark patterns are convincingly similar and would be declared a positive identification by a practicing examiner. One sit uation in which this shortcoming is especially troublesome is in poorly marked samples where striations may not be present across the entire surface of the lead plates used for making the tool marks. For example, consider the possibility where two dissimilar tools are used to mark two plates. Suppose that in both cases the screwdriver tip does not adequately mark the surface. In such cases the similar unmarked sections of the plates may produce very high correlation values, even though the marked sections are entirely dissimilar. For these and many other reasons, a simple maximized correlation coefficient is not a reliable index of match quality.

This article presents a description of a matching analysis and algorithm that overcomes many of these difficulties and summa rizes experimental data collected to characterize algorithm perfor mance. The index produced by the algorithm provides a more statistically meaningful comparison than maximized correlation. Experiments involving comparisons of samples obtained from a single tool to each other, and to samples produced from other simi lar sequentially manufactured tools, show that the analysis can fairly reliably separate sample pairs that are known matches from the same tool from pairs obtained from different tools. Addition ally, the index provides a means of calculating estimates of error rates within the narrow and specific setting of this study.

For the sake of clarity, a brief summary of how the algorithm operates and the assumptions upon which it is based is given later. This discussion is necessary to understand the algorithm results in comparison with those obtained by human subjects. Agreement between algorithm results and examiner evaluations was assessed at the 2008 Association of Firearms and Tool Mark Examiners Train ing Meeting held in Honolulu, Hawaii. Results obtained from this blind study in which practicing tool mark examiners were asked to compare the same samples will be presented. Comparison of the results obtained by human examiners to those of the algorithm provides interesting insights that will lead to algorithm performance improvements.

Statistics

An earlier work (7) described a statistical analysis and algorithm for comparing two dimensional images of tool marks. The algo rithm described here is similar in construction, although it is restricted only to matching along one dimensional profilometer data traces, and so is lacking some of the steps required to deal with two dimensional data arrays. The data examined in this analysis are of the type collected by a surface profilometer that records surface height (z) as a function of distance (x) along a linear trace taken perpendicular to the striations present in a typical tool mark. Some important assumptions in the analysis are that the values of z are reported at equal increments of distance along the trace and that the traces are taken as nearly perpendicular to the striations as pos sible. The algorithm then allows comparison of two such linear traces.

The first step taken by the algorithm, referred to as Optimization, is to identify a region of best agreement in each of the two datasets for the specified size of the comparison window (which is user defined). This is determined by the maximum correlation sta tistic, hereafter referenced as an "R value," and described in (6). By way of illustration, two different possibilities are shown in Fig. 1. The schematic of Fig. 1*a* shows the comparison of a true match, i.e., profilometer recordings from two specimens made with the same tool, while Fig. 1*b* shows data from a true nonmatch pair of specimens (i.e., two marks from two different tools). In each case, the matched regions marked with solid rectangles are the comparison windows denoting the trace segments over which the ordinary linear correlation coefficient is largest. Note that in both cases the R value returned is very close to 1, the largest numerical value a correlation coefficient can take. In the first instance this is so because a match does in fact exist, and the algorithm has suc ceeded in finding trace segments that were made by a common section of the tool surface. In the second case, the large R value is primarily a result of the very large number of correlations calcu lated in finding the best match. Even for true nonmatches, there will be short trace segments that will be very similar, and it is almost inevitable that the algorithm will find at least one pair of such segments when computing the R value. It is primarily for this reason that the R values cannot be interpreted in the same way that simple correlations are generally evaluated in most statistical settings.

For the reasons described earlier, the algorithm now conducts a second step in the comparison process called Validation. In this step a series of corresponding windows of equal size are selected at randomly chosen, but common distances from the previously identified regions of best fit. For example, a randomly determined shift of 326 pixels to the left, corresponding to the dashed rectan gles in Fig. 1*a*, might be selected. The correlation for this pair of corresponding regions is now determined. Note that this correlation must be lower than the R value, because the latter has already been determined as being the largest of all possible correlations deter mined in the Optimization step. The assumption behind the Valida tion step is that if a match truly does exist, correlations between these shifted window pairs will also be reasonably large because they will correspond to common sections of the tool surface. In other words, if a match exists at one point along the scan length (high R value), there should be fairly large correlations between corresponding pairs of windows along their entire length. However, if a high R value is found between the comparison windows of two

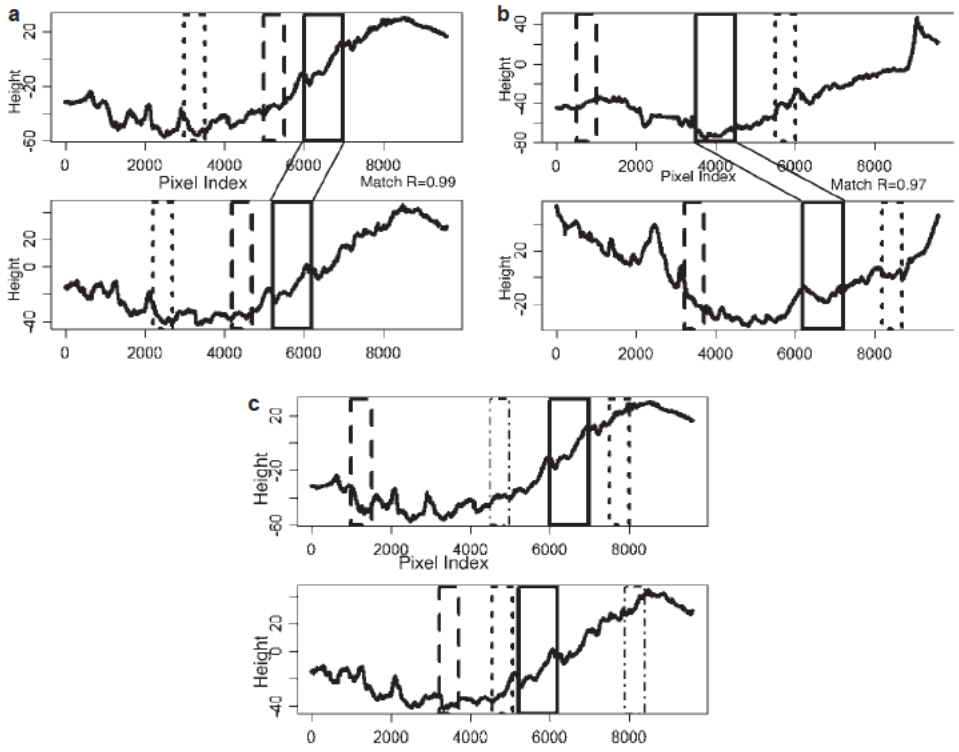

FIG. 1 *(a) Comparison pair showing a true match. Best region of fit shown in solid rectangle with corresponding* R *value. Note the similarity of the regions within the two possible sets of validation windows (dashed and dotted rectangles). (b) Comparison pair showing a true nonmatch. While a high* R *value is still found between "Match" segments, the validation windows are distinctly different from one another. (c) Validation windows (dashed, dotted, and dot and dash rectangles) selected at random for the comparison pair shown in (a) to establish a baseline value.*

nonmatch samples simply by accident, there is no reason to believe that the accidental match will hold up at other points along the scan length. In this case rigid shift pairs of windows will likely not result in especially large correlation values.

During the Validation step a fixed number of such segment pairs are identified, corresponding to a number of different randomly drawn shifts, and the correlation coefficient for each pair is com puted. Dotted and dashed rectangles displayed in Fig. 1 illustrate schematically the selection of two such pairs of shifted data seg ments; in actual operation the algorithm chooses many such pairs. In the case of the true match the regions within the corresponding dashed windows of Fig. 1*a* do appear somewhat similar and can be expected to return fairly large correlation values. However, when similar corresponding pairs of windows are taken from the non match comparison of Fig. 1*b*, the shape of the scans within the windows is seen to differ drastically. Lower correlation values will be obtained in this case.

The correlation values computed from these segment pairs can be judged to be "large" or "small" only if a baseline can be estab lished for each of the sample comparisons. This is achieved by identifying a second set of paired windows (i.e., data segments), again randomly selected along the length of each trace, but in this case, without the constraint that they represent equal rigid shifts from their respective regions of best fit. In other words, for this second set of comparisons the shifts are selected at random and independently from each other any segment of the selected length from one specimen has an equal probability of being compared to any segment from the other. This is illustrated in Fig. 1*c* for three pairs of windows, denoted by the dashed rectangles, the dotted rect angles, and the dot and dash rectangles.

The Validation step concludes with a comparison of the two sets of correlation values just described, one set from windows of common random rigid shifts from their respective regions of best agreement, and one set from the independently selected windows. If the assumption of similarity between corresponding points for a match is true, the correlation values of the first set of windows should tend to be larger than those in the second. In other words, the rigid shift window pairs should result in higher correlation val ues than the independently selected, totally random pairs. In the case of a nonmatch, as the identification of a region of best agree ment is simply a random event and there truly is no similarity between corresponding points along the trace, the correlations in the two comparison sets should be very similar.

A nonparametric Mann Whitney *U* statistic (referred to in this article as T1), computed from the joint ranks of all correlations computed from both samples, is generated for the comparison. Where the correlation values of the two comparison sets are simi lar, T1 takes values near zero, supporting a null hypothesis of "no match." If the correlations from the first rigid shift sample are sys tematically larger than the independently selected shifts, the result ing values of T1 are larger, supporting an alternative hypothesis of "match."

Method

The test set for this study is the same as described in (6), namely, a series of 50 sequentially manufactured screwdriver tips were obtained and used to make tool marks at angles of 30°, 60°, and 85° on flat lead plates. The surface roughness of the resultant striae was measured using a surface profilometer and the measure ments saved as a series of data files detailing *z* height as a function of *x* direction. All details of data collection are given in (6).

To compare the effectiveness of the algorithm to human examin ers, and potentially identify areas where the algorithm might be

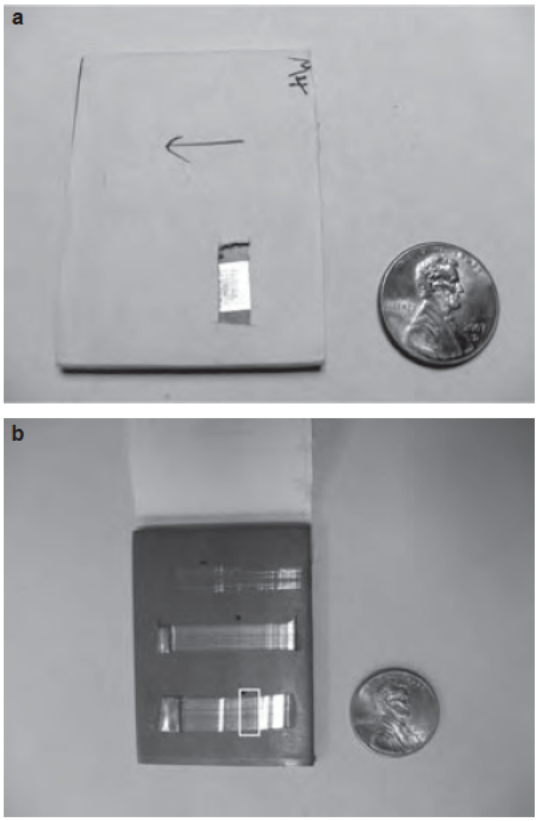

FIG. 2 *Image of a tool marked plate with (a) a blinder in place; (b) the blinder removed, a white rectangle shows the visible portion when the blinder is in place.*

enhanced or improved, a double blind study was conducted during the 2008 Association of Forearms and Tool Mark Examiners Training Seminar. During the course of this meeting, 50 different volunteers rendered over 250 opinions on some of the sample pairs used for this study and evaluated by the algorithm.

A series of 20 comparison pairs covering a range of T1 values from low to high were selected from the tool marks produced at the 85° comparison angle. Of the 20 comparison pairs, five were from samples where the algorithm correctly identified a matched set (high T1); five were correctly eliminated nonmatch comparisons (low T1); five were incorrectly eliminated matched sets (T1 values in the low or inconclusive range); and five were incorrectly identified nonmatches (intermediate or high T1). Examiners were asked to assess each pair of samples twice. For the initial observation, paper blinders were placed on the samples so that examiners were restricted in their view to the same general area where the profilometer data were collected (Fig. 2). After making an initial assessment, the blinders were removed and the examiner was given the opportunity to make a second assessment based on a view of the entire sample. In each case, examiners were asked to render an opinion as to whether they were viewing a positive identification, a positive elimination, or inconclusive, for reasons that will become apparent.

Names of examiners were not recorded, although demographic data were collected concerning the experience and training of the volunteers. Of the 50 volunteers, all except five were court qualified firearm and tool mark examiners. Of the remaining five, two were firearms (but not tool mark) qualified, two were in training, and one was a foreign national where a court qualification rating does not exist. Volunteers were required to do a minimum of two comparison pairs and could do as many as they wished. Several chose to do the maximum number of comparisons possible. Numbers were assigned to identify each volunteer during data collection; afterward the ID numbers were randomly mixed to preserve anonymity.

Examiners were asked to use whatever methodology they employed in their respective laboratories. This caused some confusion initially and placed constraints on the volunteers because some laboratories never use the term "positive elimination," while others are reluctant to use the term "positive identification" unless the examiner personally either makes the marks or knows more information about them than what could be supplied in this study. After understanding this, the examiners were told the direction of the tool when making the mark and that the tool marks were all made at the same angle from similar, sequentially made, flat blade screwdriver tips. Also, examiners were told that for the study they could consider the terms of "positive elimination" or "inconclusive" to be essentially interchangeable.

Results and Discussion

Algorithm Performance

The data obtained from the profilometer were used to test a series of hypotheses that are held as being true by tool mark examiners (Fig. 3). The first and most fundamental assumption, that all tool marks are unique, was tested by a comparison of marks made by different screwdriver tips at the angles of 30°, 60°, and 85° with respect to horizontal. The T1 statistic values are shown in Fig. 4 as a function of angular comparison. The data are plotted as box plots, the boxes indicating where 50% of the data falls with the spread of the outlying 25% at each end of the distribution shown as dashed lines. As stated previously, when using a T1 statistic a value relatively close to 0 indicates that there is essentially no evidence in the data to support a relationship between markings. For pairs of samples made with different screwdrivers (Fig. 4) the majority of the index T1 values produced by the algorithm fall near the 0 value; only three outlier comparisons had a T1 value greater than ±4.

In comparison, Fig. 5 displays indices computed using the algorithm from profilometer scans of marks made by the same side of the same tool and compared as a function of angle. While marks made at different angles still produce index values near 0, the T1 statistic jumps dramatically when marks made at similar angles are considered. Clear separation is seen between the 50% boxes, although overlap still exists when the outliers are considered.

Taken together, Figs. 4 and 5 support Hypotheses 1 and 2. When comparing tool marks made at similar angles with different tools, the resulting T1 values cluster near zero (Fig. 4), but when the same tool is used to make marks at similar angles, the T1 distributions are on substantially larger values, giving support for Hypothesis 1. Support for Hypothesis 2 is demonstrated by Fig. 5 alone, because even among the same tool marks, only those made at the same angle produce large T1 values.

Hypothesis 1:	*The 50 sequentially produced screwdrivers examined in this study all produce uniquely identifiable tool marks*
Hypothesis 2:	*In order to be identifiable, tool marks from an individual screwdriver must be compared at similar angles.*
Hypothesis 3:	*Different sides of a flat-bladed screwdriver produce different uniquely identifiable marks.*

FIG. 3 *Summary of hypothesis tested in this study.*

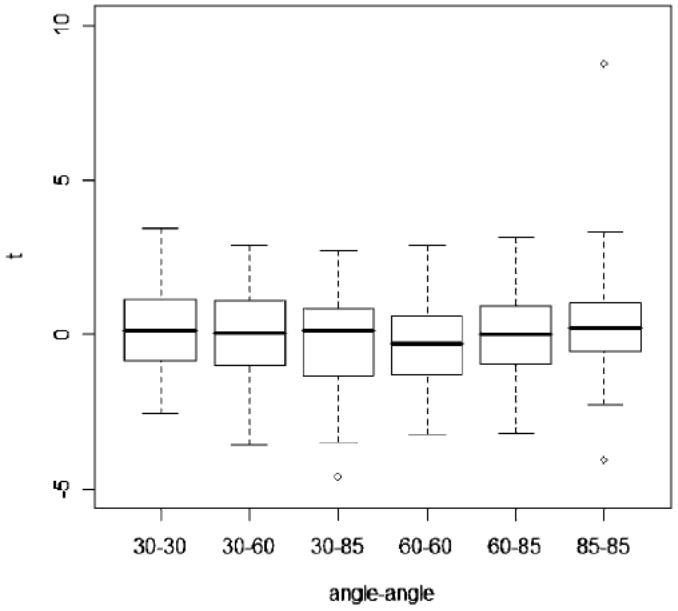

FIG. 4 *Box plots showing T1 results when comparing marks from different screwdrivers.*

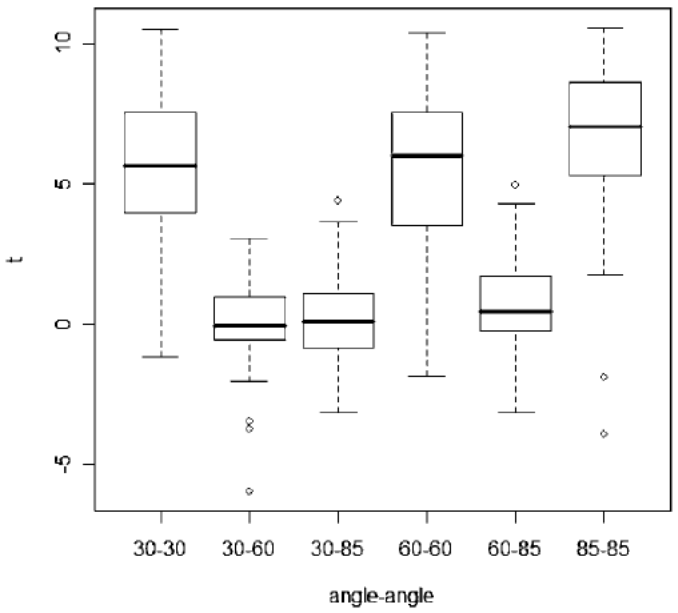

FIG. 5 *Box plots showing T1 results when comparing marks obtained from the same side of the same screwdrivers.*

The last hypothesis considered was that when comparing tool marks made from screwdriver tips, the marks must be made from the same side of the screwdriver; marks made using different sides of the screwdriver appear as if they have come from two different screwdrivers. These results are shown in Fig. 6. The hypothesis is again supported because, as in Fig. 4, the T1 values cluster around 0 regardless of the angles used in making the marks, indicating no relationship between the samples.

The T1 values summarized in Figs. 4 and 5 are individually re plotted in Fig. 7, with the *y* axis randomly varied (known as jitter ing) to create an artificial vertical separation that makes it easier to view the data points. Known comparisons that should match and

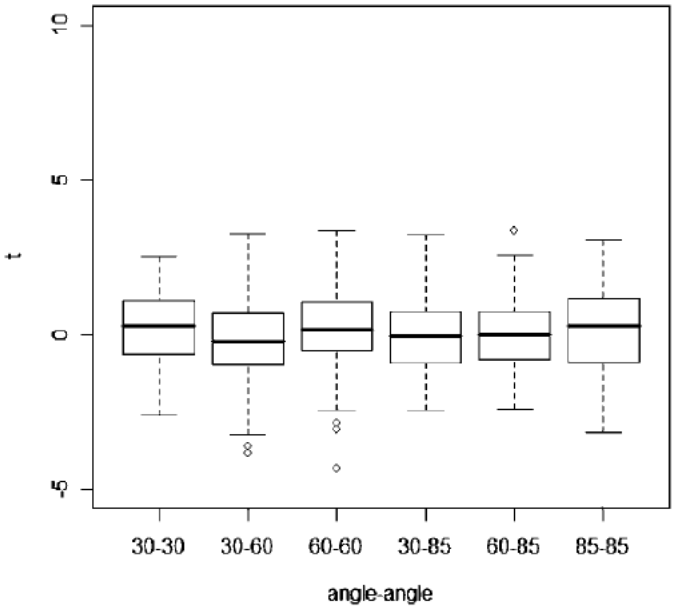

FIG. 6 *Box plots showing T1 results when comparing marks made from different sides of the same screwdrivers.*

produce high T1 values are shown in black. Known "nonmatches" that should have T1 values near zero are shown in gray.

Examination of these plots indicates that the algorithm operates best using data obtained at higher angles than lower angles, i.e., the spread of black and gray spots is more defined for the 85° data than, for example, the 30° data. This is believed related to the qual ity of the mark. As the angle of attack of the screwdriver with the plate increased, the quality of the mark increased. It was common to obtain marks that represented the entire screwdriver tip at high angles, while marks at lower angles were often incomplete (5). Algorithm performance also appears more efficient at reducing false positives than it does in eliminating false negatives. At all angles known matches were found with very low T1 values, while nonmatches with high T1 values were very limited.

While T1 is a much more stable index of match quality than *R* value, problems still remain in establishing an effective, objective standard for separating true matches from nonmatches. Ideally, when employing standard *U* statistic theory the critical T1 values separating the regions of known matches (black data points) and known nonmatches (gray data points) should remain constant for all datasets. Examination of Fig. 7 shows that this is not the case. For example, reasonable separation for the 30° and 60° data appears to be somewhere around a T1 value <5, but rises to approximately 7 for the 85° data. This variation is most likely attributed to the lack of complete independence among the correla tions computed in each sample, arising from the finite length of each profilometer trace.

For the reasons discussed earlier, assigned threshold values indi cating "Positive ID" and "Positive Elimination," and denoted by black lines on the graphs of Fig. 7, were chosen based on a *K* fold cross validation using 95% one sided Bayes credible intervals. Specifically, the lower threshold is a lower 95% bound on the 5th percentile of T1 values associated with nonmatching specimen pairs, and the upper threshold is an upper 95% bound on the 95th percentile of T1 values associated with matching specimen pairs. The region between these two threshold values is labeled "Incon clusive." A Markov Chain Monte Carlo simulation was used to determine potential error rates.

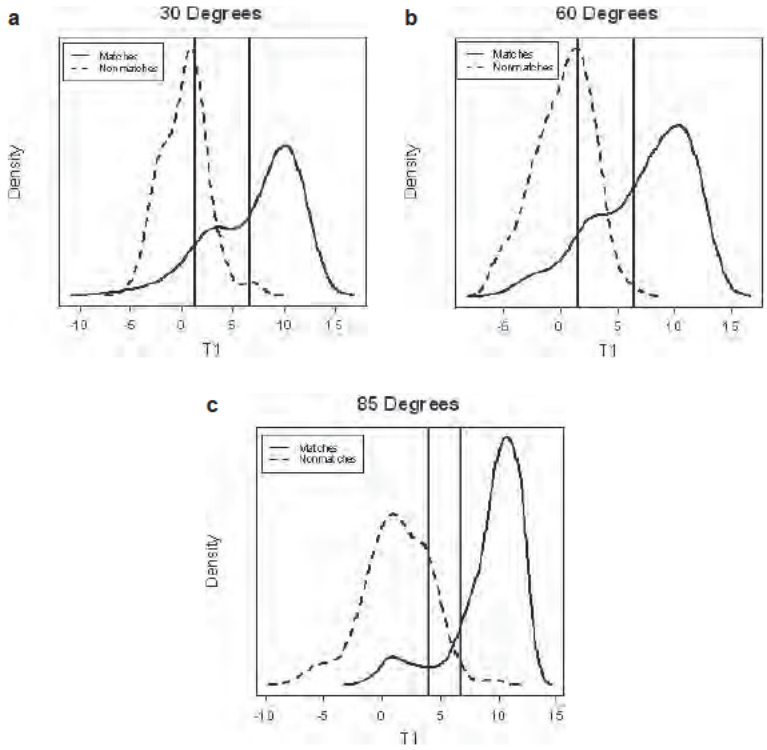

FIG. 7 *Summation of the T1 values from comparisons made at (a) 30°; (b) 60°; (c) 85°.*

Using this method the estimated error rates are as follows. For comparisons made at 30° the estimated probability of a false positive (i.e., a high T1 value for a known nonmatch comparison) is 0.023. In other words there is a possibility of slightly over two false positives for approximately every 100 comparisons. The estimated probability of a false negative is 0.089, or almost nine true matches having a low T1 value per every 100 comparisons. The cross validation method used ensures that all the data have similar error rates, and the rates found for the 60° and 85° data are approximately 0.01 and 0.09 for false positives and false negatives, respectively. What is most noticeable is that the T1 lower threshold value for the 85° data is much larger than for the 30° and 60° data, being 4.10 versus 1.34 and 1.51, respectively. This suggests that a more distinct difference is required to classify nonmatches for the 30° and 60° cases than is true for the 85° case. This, in turn, results in a corresponding increase for the estimated inconclusive error rates, which are 0.103, 0.298, and 0.295 for the 85°, 60°, and 30° data, respectively. It would, of course, be possible to shift these error rates, i.e., produce fewer false negatives at the expense of more false positives, by altering the percentiles used in our estimation procedure.

Association of Firearm and Tool Mark Examiners Study

Results of the computerized analysis of specimen pairs were compared to expert evaluations of the same samples made by volunteer examiners at the 2008 Association of Firearm and Tool Mark Examiners seminar. However, before the algorithm performance can be discussed in comparison with the data obtained at the Association of Firearm and Tool Mark Examiners seminar using human volunteers, a brief consideration of the constraints experienced by the examiners is in order. First, it should be recognized that the conditions under which the examiners rendered an opinion would ordinarily be regarded as restrictive or even professionally unacceptable. Without having the tool in hand, or without being permitted to make the actual mark for comparison, tool mark examiners were forced to make assumptions they would not make in an actual investigation. For example, without having the screwdriver tip in hand the examiners did not know whether the mark they observed represented the entire width or only a portion of the screwdriver blade. Second, given this uncertainty about how the specimen was made, examiners tended to be more conservative in their willingness to declare a positive identification or elimination. During the course of the Association of Firearm and Tool Mark Examiners study, several examiners commented that typical lab protocol would require them to have physical access to the subject tool before rendering a "positive identification" judgment. Finally, examiners do not typically employ the terms used to denote the three regions identified for error analysis. Thus, while privately saying they felt a comparison was a "positive elimination" (given their knowledge of the test being conducted), lab protocol required an opinion of "inconclusive" to be rendered. Such policies are put in place because the signature of a tool may so change during use that a mark made at one point in time may not resemble a mark made with the same tool at a different point in time, e.g., after the tip has been broken and/or reground. In such cases positive elimination is only allowed if the class characteristics of the marks are different.

When viewed in light of these constraints, some interesting observations concerning the algorithm performance are apparent. In a small number of cases (12 out of 252 comparisons), when examining the entire tool mark after first viewing only the restricted area where the profilometer scans were obtained, examiners changed their opinion from inconclusive to either positive ID or positive elimination. This indicates that algorithm performance might be improved simply by increasing the amount of data processed. This may be achieved, for example, by ensuring that the profilometer scans span the entire width of the mark or possibly by considering

a number of scans taken at locations dispersed along the entire length of the available mark.

In a slightly smaller number of cases, comparisons between specimens made by the same screwdriver that were not conclu sively identified as such by the algorithm also presented problems for the examiners. Five true matches that received low T1 values and were classified as a positive elimination by the algorithm were examined during the Association of Firearm and Tool Mark Exam iners study. Three of the five were given ratings of "inconclusive" or "positive elimination" on one occasion, and one particular com parison sample (designated MW4) was rated this way seven times. Thus, while examiners in general were vastly superior to the algo rithm in picking out the matches, both the algorithm and the exam iners had more trouble with some true matches than with others.

Close examination of the sample that was most often problem atic for examiners (i.e., MW4) was conducted, and the images obtained are shown in Fig. 8. Figure 8*a* shows the side by side comparison of the marks, where no match is seen. Note that the mark width matches extremely well, and the entire mark seems to be present. Figure 8*b* shows the samples positioned where the true match is evident. It can be seen that each mark only represents a portion of the screwdriver blade width, predominantly from the two sides of the tip. A match is only possible if the marks are offset, allowing the opposing "edge" sections (which actually were produced by the middle of the screwdriver blade) to be viewed side by side.

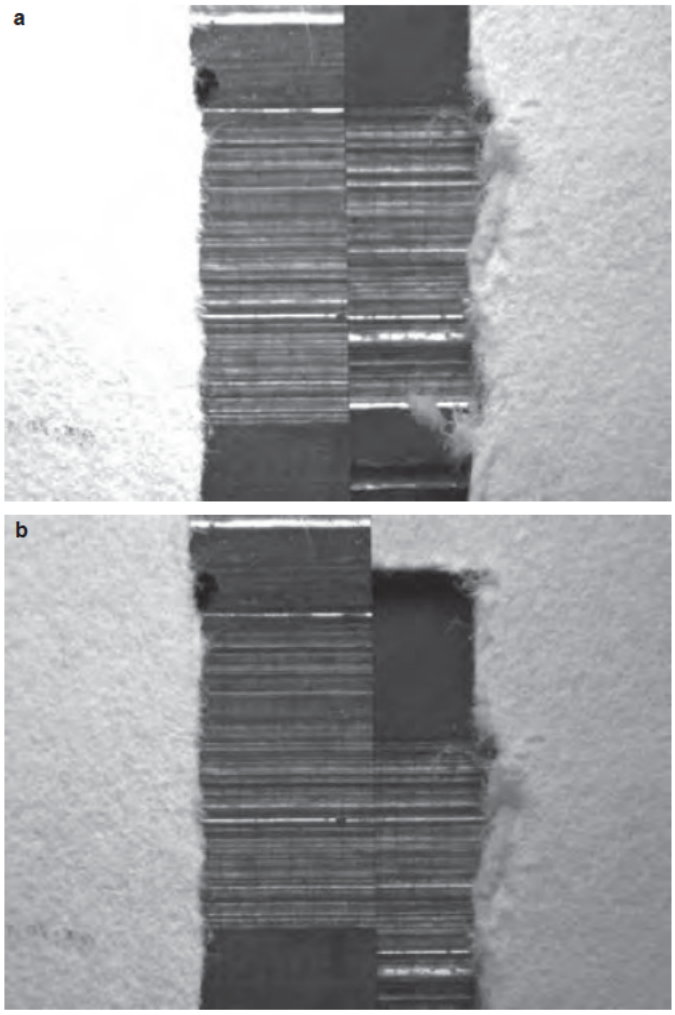

FIG. 8 *Sample MW4 with (a) tool marks placed so that assumed edges align; (b) correct placement required for positive identification.*

This sample points out weaknesses in the study conducted at the Association of Firearm and Tool Mark Examiners as well as in the laboratory tests of the algorithm. In a screwdriver mark comparison it is common for examiners to use the edges of the marks as initial registration points for the start of an examination. As examiners make the comparison marks themselves they are well aware of the edge markings, if not for the evidence marks, at least for the marks they produced. In the Association of Firearm and Tool Mark Examiners study, such information was not provided and may have led to some false assumptions. For example, in the majority of cases the volunteers were under some pressure to quickly conduct a comparison before, e.g., the next meeting session started, or so that another examiner could use the equipment, etc. Because of these time constraints, samples often were placed on the stages of the comparison microscope for the volunteer, giving the examiner little or no time to observe the macroscopic appearance of the mark. Without the benefit of seeing the size of the entire mark, and given the identical widths of the two partial marks for sample MW4 when initially viewed using the comparison microscope, the assumption that the entire width of the screwdriver blade was represented would be a natural one. However, such an assumption could easily lead to an inconclusive or positive elimination conclu sion, especially if the examiner was being conservative because of the lack of information concerning the sample.

The problem described earlier essentially relates to the examiners having a lack of a point of reference or registry of the mark for the comparison. The same could be said of the algorithm and the man ner in which it performs, because no point of registry exists to indi cate when the data being acquired is actually coming from a tool marked region or from the unmarked plate. All of the profilometer scans analyzed by the algorithm were run using the same set of sampling parameters. However, the initial positioning of the stylus was inexact. For incomplete marks, large regions of the unaffected lead plate were also scanned to keep the file sizes consistent, and this lack of registry could have affected algorithm performance. This is not immediately evident if one examines the raw profilome ter traces (Fig. 9). In this figure the top and bottom traces show the entire scans while the two middle traces show the matched details found within the two corresponding solid rectangles superimposed on the top and bottom traces. At first sight the two scans do appear quite different, as the offset in the scans, revealed during examina tion at the Association of Firearm and Tool Mark Examiners, is not immediately evident in the data files. Given observation of Fig. 8, one can mark the approximate location of the region that is common between the two traces; this is shown in Fig. 9 by the shaded rectangles. In this case, paired validation windows displaced equal amounts in either direction may return a low T1 value because the majority of either scan is not held in common with the other. In other words, there is a high probability that the validation windows fall in regions where no correspondence between plates exists (see Fig. 8*b*). Thus, what should be a match is rated as a nonmatch.

A somewhat different problem is revealed when traces from true nonmatch samples are examined (Fig. 10). In these instances, the optimization step may identify windows at extreme edges of the two traces as being most similar. Given the nearness of the match to the ends of the traces, the random selection of paired, rigid shift windows during the validation step is severely constrained. For the example shown in Fig. 10*a*, the match region (denoted by solid rectangles) falls at the extreme right ends of the data files. This means that the rigid translations taken for each pair of verification

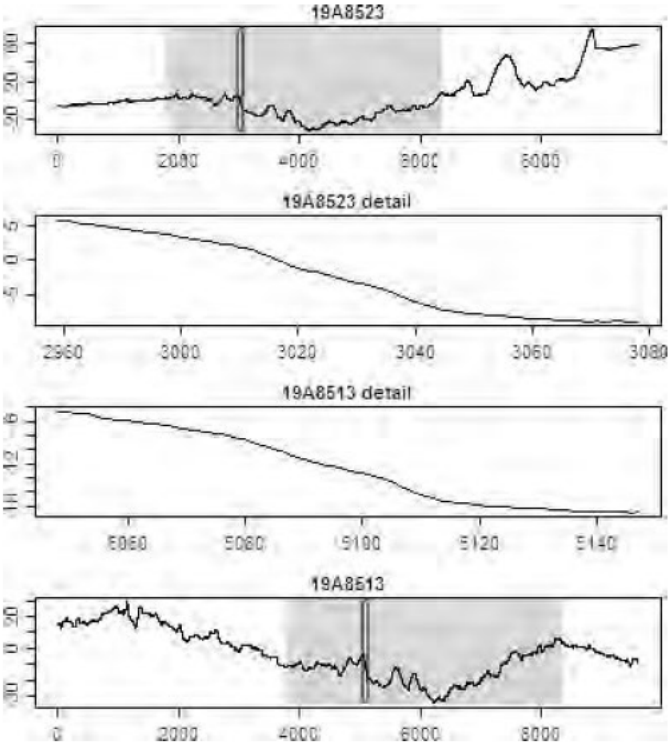

FIG. 9 *Profilometer data showing results from comparison MW4. The match region is shown by the solid rectangle.*

windows must always fall to the left of the match region. While this may be less than desirable, the entire mark is still available for validation and a large number of rigid shift windows spaced across the entire length of the file should be sufficient to produce good separation between this accidental match and the T1 values of a true match. However, this is not true for the true nonmatch shown in Fig. 10*b*. In this case the windows identified in the optimization step as being most similar are at opposite ends of the compared data traces. The distances of possible rigid translations are con strained to a short distance to the left of the top profile and a short distance to the right of the bottom profile. Thus, the majority of the mark cannot be used in the validation step for this accidental match. If the regions in the immediate vicinity of the accidental match are also similar, high T1 values may be returned because of the constrained sampling parameters, giving results that cannot be separated from a true match.

The earlier discussion suggests that further development of the algorithm to incorporate additional data concerning the region of the profilometer trace that is actually tool marked and/or the loca tion of the tool edge might improve its performance. While tool mark examiners do not directly use features such as these as a basis for identification, they do use it indirectly in establishing a context for the comparison. Such information, routinely and quickly noted by a human examiner, is unavailable to the current algorithm. The algorithm treats all possible pairs of trace windows the same way and functions under the assumption that all marks analyzed are essentially the same, i.e., it assumes the screwdriver tip has completely marked the lead plate and that no unmarked regions exist. This clearly is not the case. At this time it appears the best way to enhance algorithm performance is to ensure that all compar ison windows (i.e., Match and Validation) are taken from regions representing the true marked surface of the lead, and that most sim ilar windows found at the trace edges are used as a basis for match

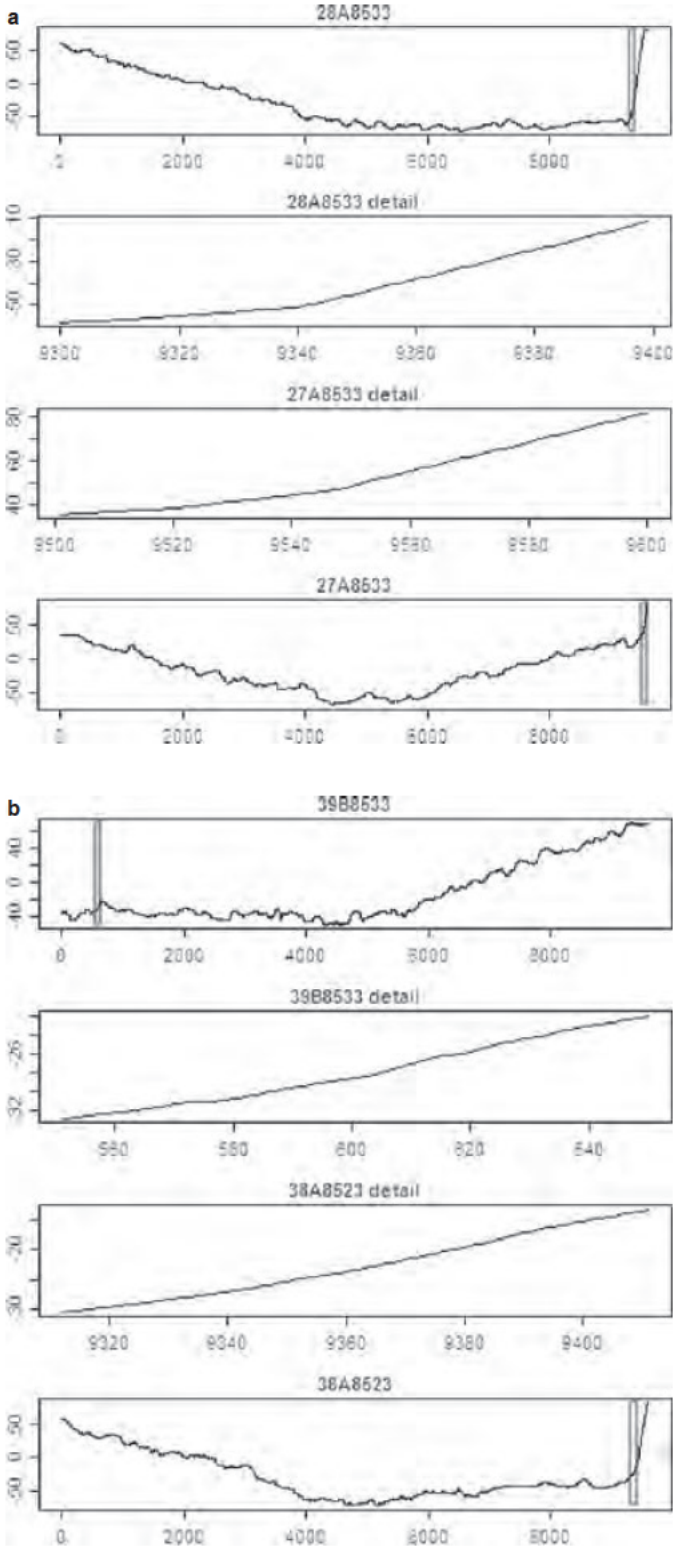

FIG. 10 *Comparison of traces obtained from four different screwdrivers that were rated as possible matches by the algorithm. (a) Good agreement found at similar ends of the traces resulted in high T1 numbers for known nonmatching pairs. (b) Good agreement found at opposite ends of the traces resulted in high T1 numbers for known nonmatches.*

identification only if they are found at the same end of their respec tive traces.

As a final comment, it should be noted that all types of volun teers (practicing examiners, trainees, retired examiners) were

involved in the study, with records kept as to the experience of the participant. Examination of the demographic data in relation to the results showed no significant difference between experienced examiners and rather newly qualified examiners or those in training; all performed equally well.

Summary and Conclusions

The analysis described here for comparing two tool marked plates is a substantial improvement over simply identifying regions of highest correlation. It does this by producing a nonparametric Mann Whitney statistic, here called T1, obtained through an optimization step followed by a validation step as a measure of evidence for tool mark matching. When used in evaluating the three hypotheses tested, namely, the uniqueness of tool marks, the necessity of comparing marks at similar angles, and the uniqueness of different sides of screwdriver blades, the T1 statistic results constitute support for the experiential knowledge of tool mark examiners. Analysis of algorithm performance in light of actual examiner results reveals deficiencies in algorithm performance that can now be addressed. Increasing the data input, possibly by including more scans spread over a larger tool mark area and incorporating contextual information normally available to examiners, such as the presence of partial scans and reference points from tool edges, may lead to performance improvements. Such changes should, for example, prohibit the identification of opposite end windows in the optimization step as potential match segments.

It is clear that examiner performance was much better than the algorithm. While the 20 samples examined at the Association of Firearm and Tool Mark Examiners represent only a subset of the total comparisons examined using the algorithm, they did contain those samples that were most definitively misclassified by the algorithm. For example, of the 20 true match pairs shown to the Association of Firearm and Tool Mark Examiners volunteers, the algorithm correctly identified 10 of the 20 samples unambiguously; the remaining 10 were listed either as inconclusive or misidentified as a false negative. In comparison, only 11 out of 126 volunteer examinations resulted in false negative classification of true match pairs, primarily from sample MW4 (7 out of 11). Further, the Association of Firearm and Tool Mark Examiners volunteers reported no false positives at all. (N.B. The caveat must be added that the terminology used in the previous statements regarding errors is not entirely consistent with examiner protocols and should not be construed by the reader to suggest that the examiners erred. Examiners are trained to render an opinion of positive identification only when no doubt exists in their minds. Thus, a false negative only suggests that the examiner was not fully persuaded.)

Acknowledgments

The authors are extremely grateful to Wayne Buttermore of Leica Microsystems and Kevin Boulay and Mike Howell from Leeds Precison Instruments for providing comparison microscopes. Without their assistance much of this study could not have been conducted. We are also grateful to officers and organizers of the 2008 Association of Firearms and Tool Mark Examiners Training Seminar held in Honolulu, especially Jim Hamby, Cindy Saito, and Curtis Kubo, for helping us with the booth and getting volunteers for the study. Finally, we gratefully acknowledge the assistance of all the AFTE members who took the time to participate in our study. This study was supported by the National Institute of Justice under contract 2004 I J R 088, and was performed in part at Ames Laboratory, which is operated under contract No. W 7405 Eng 82 by Iowa State University with the US Department of Energy.

References

1. Bisotti A. A statistical study of the individual characteristics of fired bullets. J Forensic Sci 1959;4(1):34–50.
2. Bonfanti MS, DeKinder J. The influence of the use of firearms on their characteristic marks. AFTE Journal 1999;31(3):318–23.
3. Bonafanti MS, De Kinder J. The influence of manufacturing processes on the identification of bullets and cartridge cases – a review of the literature. Sci Justice 1999;39:3–10.
4. Bisotti A, Murdock J. Criteria for identification or state of the art of firearm and tool mark identification. AFTE Journal 1984;16(4):16–24.
5. Committee to Assess the Feasibility, Accuracy and Technical Capability of a National Ballistics Database, National Research Council for the National Academy of Sciences. Ballistic imaging. Cork DL, Rolph JE, Meieran ES, Petrie CV, editors. Washington, DC: National Academies Press, 2008.
6. Faden D, Kidd J, Craft J, Chumbley LS, Morris M, Genalo L, et al. Statistical confirmation of empirical observations concerning tool mark striae. AFTE Journal 2007;39(3):205–14.
7. Baldwin D, Morris M, Bajic S, Zhou Z, Kreiser MJ. Statistical tools for forensic analysis of tool marks. Ames (IA): Ames Laboratory Technical Report, 2004; IS 5160, http://www.osti.gov/bridge/product.biblio.jsp?osti_id=825030.

Additional information and reprint requests:
L. Scott Chumbley, Ph.D.
Materials Science and Engineering Department
Iowa State University
2220 Hoover
Ames, IA 50011
E mail: chumbley@iastate.edu

Cartridge Case and Bullet Comparison Validation Study with Firearms Submitted in Casework

By: Erich D. Smith, Firearms/Toolmarks Examiner, Federal Bureau of Investigation, Quantico, VA

Key Words: Validation, cartridge case, bullet, identification, comparison, class characteristics, individual characteristics, blind test, pattern recognition, manufacturing marks, forensic firearms identification, accuracy, precision

ABSTRACT

This validation study was designed to test the accuracy and the reproducibility (precision) of examinations performed by trained firearms examiners who use pattern recognition for identification.

Introduction

The theory underlying firearms identification is that no two firearms should produce the same microscopic features on bullets and cartridge cases such that they could be falsely identified as having been fired from the same firearm. This is possible because the microscopic features produced on the surface of bullets and cartridge cases are a direct result of the following variables for barrels and breech faces: tool(s) used for manufacture and their state of wear, usage of the barrel and breech, environmental exposure and abuse. These variables are know to produce random microscopic contours on the surfaces of barrels and breech faces. Patterns produced on bullets and cartridge cases from contact with these surfaces, can be microscopically compared to determine if they have originated from a common source.

In the field of forensic firearms identification, pattern recognition enables examiners to make identifications. Pattern recognition, for firearms examiners, is the ability to individualize firearms through physical comparison of the microscopic marks on bullets and cartridge cases. This is possible because many humans have the ability to recognize degrees of correspondence in patterns. This ability to recognize patterns, combined with specialized training in firearms identification, makes the discipline of forensic firearms identification possible.

The ability of an examiner to determine an identification requires training and an understanding of the individualizing patterns produced by firearms. Before a microscopic comparison begins, a foundation is built by measuring and comparing available class characteristics, such as General Rifling Characteristics (GRCs). These objective criteria are used to narrow the pool of candidates for determining a common source. Once an available foundation has been established, a common source often can be determined by evaluating individual microscopic marks of value using pattern recognition.

During training, an examiner begins to develop an identification threshold - a subjective point where sufficient agreement in the individual microscopic marks of value can determine a common source. An examiner's identification threshold develops by his examining numerous known matches and known non-matches to understand what is necessary for sufficient agreement for identification. Despite the goal for the threshold which examiners develop during training to be consistent among qualified examiners; the subjective point for an identification for qualified examiners may not be equal for any given group of examiners, and over time, the threshold of an individual examiner may change.

In court proceedings firearms examiners are sometimes asked whether two barrels or breech faces could produce enough similarity in microscopic marks such that bullets and cartridge cases could be incorrectly identified as having been fired from the same firearm. This question is a two-fold challenge to firearms identification – whether firearms identification can determine a common source through pattern recognition of the individual microscopic marks (accuracy); and, whether examiners as a whole are proficient enough to obtain the same results for a particular examination to determine a common source (precision).

Prior validation studies have supported the underlying theory of firearms identification by obtaining test samples from consecutively manufactured barrels. (1,2) These studies tested firearms identification because they considered the likelihood that two barrels would have microscopic similarities in/on their interior surfaces, due to minimal tool wear from consecutively

Date Received: December 8, 2003
Peer Review Completed: May 4, 2005

manufactured barrels. In these studies no misidentifications were recorded using consecutively manufactured barrels.

This study departs from the earlier studies by examining the variables that generate the microscopic features on barrels and breech faces over an extended period of time. Firearms with similar class characteristics and similar manufacturing techniques were selected. The selected firearms circulated in the general population where they were exposed to the environment and abuse.

Further, this validation study sought to challenge preconceptions that examiners might have developed from taking earlier proficiency or validation tests. Participants in proficiency or validation tests may or may not consciously anticipate particular answers to questions. This can occur when an examiner recognizes a particular test pattern or design or expects a series of identifications or exclusions to be included in the test. Possible preconceptions to test answers or test design were challenged by presenting test samples that did not produce an easily expected or predictable result.

Test Design

This test featured bullets and cartridge cases from firearms submitted from casework, such specimens having similar class and individual microscopic marks. However, only two matches were present out of seven hundred and twenty comparisons. The overwhelming number of non-matches sought to challenge the examiners' identification threshold using pattern recognition while challenging testing preconceptions using test specimens that could be experienced during actual casework.

Each test packet contained one true identification and forty-four true eliminations (exclusions) for both cartridge case and bullet comparisons. This created a total of three hundred and sixty comparisons for the eight examiners with eight true identifications and three hundred and fifty-two eliminations for both cartridge cases and bullets.
Eight firearms examiners from the FBI Laboratory's Firearms/Toolmark Unit (FTU) participated in this study. The participants ranged from the recently trained to an examiner with twenty years experience. The participants were instructed to conduct their own examinations as if they were performing an examination on assigned casework.

Set-up

During a four-month period, nine Ruger P89 pistols were collected in the FBI's Laboratory. The P89 was selected as a test specimen because of its availability and the manufacturing techniques used to produce the barrel and breech. The nine pistols were delivered to the FBI's Laboratory from various FBI field offices across the United States. Seven pistols were collected upon receipt into the Laboratory. Two pistols were randomly selected from the FBI Laboratory's Reference Firearm Collection (RFC). The cities of origin were determined through their case number and serial numbers were used to determine the approximate date of manufacture. (Table 1)

Remington UMC brand cartridges were selected for test firing. The same lot of ammunition was used for all test fires. The cartridges consisted of 115 Grain, copper-jacketed bullets with open base, brass cases, and nickel primers. Nine firearms were test fired separately in the FBI Laboratory's water tank ten times. One of the nine firearms was test fired an additional ten times for a total of twenty test fires for that single firearm. Before and after each test fire the serial numbers were confirmed and recorded. After each test fire the specimens were placed into a container marked with the serial number of the firearm. Each bullet and cartridge case was assigned a unique identifier consisting of a letter and two numbers corresponding to the firearm from which they were test fired. The GRCs of each firearm were measured and recorded. The bullets and cartridge cases were examined using a comparison microscope to confirm that individual microscopic marks of value were reproduced between test fires. The marked, test fired specimens were separated into nine containers, each having eight cartridge cases and eight bullets from different firearms and two bullets and cartridge cases from the same firearm. The containers were marked with a code consisting of a letter and three numbers, and then sealed.

Nine test packets were prepared. Each packet contained ten cartridge cases, ten bullets, a comparison sheet, and directions for performing the validation study. Each packet was given its own identifier to maintain the anonymity of the test participant. Each participant received a packet randomly. The comparison sheet had the corresponding code for the container, a list of comparison combinations, and space to record a result. The examiners were instructed to use the following nomenclature for their answers: (I) Identification, (NC) No conclusion, (NI) Non-identification (an exclusion).

At the completion of the test each examiner was required to return all materials, notes and handouts to the proctor's mailbox. Each participant was instructed not to discuss any of their findings with the other participants.

Results

Eight participants received test packets and eight test packets were returned completed. There were a total of three hundred and sixty cartridge case comparisons with no false positives (an incorrect association) and no false negatives (an incorrect

non-identification). The results for cartridge cases recorded by the examiners were seven identifications, three hundred and thirty-five no conclusions, and eighteen non-identifications. (Table 2)

The majority of cartridge cases comparisons were true non-identifications having similar class characteristics. This design produced a dilemma for the participants' identification threshold. Some participants may have expected a larger number of comparisons to be identifications. However, the results indicate the majority of examiners were not misled by the overwhelming number of non-identifications and were able to reach correct reach conclusions, with no false positives.

One participant recorded a no conclusion result for a true identification – this is acceptable and scientifically sound, indicating an insufficient agreement of individual microscopic marks of value to formulate the identification. The lack of agreement can be a result of external variables such as varying pressures between test fires, wear in the microscopic marks, environment, abuse and debris creating ambiguity in the individual microscopic marks from consecutive test fires from a single firearm. Internal variables such as the examiner's experience level, test anxiety and test design may have resulted in heightened conservatism.

The large number of no conclusions recorded for the true non-identifications were expected for this validation study. The fundamental elimination criterion utilized by FBI examiners requires a difference in class characteristics to reach an elimination conclusion with comparison specimens. External variables, described earlier, can influence the microscopic detail, especially regarding bullets. This class difference criterion reduces the possibility of a false elimination.

One participant was able to correctly eliminate eighteen cartridge cases. As noted above, the criteria for elimination by the FTU requires a difference in class characteristics. The firing pin apertures did exhibit a measurable design difference in their sizes. The previously described external and internal variables would have contributed to the other participants choosing an inconclusive response with these test specimens.

There were a total of three hundred and sixty bullet comparisons. None of these involved false positives or false negatives. Each packet contained one true identification and forty-four true eliminations.

Again, the majority of recorded comparisons were "no conclusion" results. This was an expected outcome because GRCs between all the bullets were similar. Adhering to their elimination criterion on class characteristics, the examiners were left to their own identification threshold to determine if an identification existed between comparisons. The results indicate that the majority of examiners were not misled by the overwhelming true eliminations and were able to determine sufficient agreement in the individual microscopic marks for identifications. Two examiners recorded a no conclusion result, which is not an incorrect response. The two no conclusion results can be attributed to the external and internal variables. (Table 3)

Conclusion

Although this test was designed to mimic actual casework, this cannot be achieved entirely. The examiners understood they were participating in a blind validation study, and that an incorrect response could adversely affect the theory of firearms identification. This creates a potential bias on the part of the participants to be conservative when answering difficult comparisons using pattern recognition. This potential bias could result in fewer identifications. However, the results of this study reflect the contrary. The majority of participants were able to determine the true identifications amongst the overwhelming number of true eliminations. The results indicate that the participants' comparisons were precise, using pattern recognition to determine a common source.

In addition, the absence of false positives or false negatives indicates that the theory of firearms identification, using pattern recognition, is an accurate and precise method for determining a common source for bullets and cartridge case for firearms collected from casework.

Acknowledgments

This author would like to thank the following for their assistance: Unit Chief Stephen G. Bunch Ph.D., Physical Science Technicians Timothy D. Zema, Albert S. Rollins, and Tammy Mullen, Administrator, Armourer School, Sturm, Ruger & Co., Inc. for her assistance in collecting information on the selected firearms.

Reference:

1.Brundage, David J. "The Identification of Consecutively Rifled Gun Barrels." AFTE *Journal*, vol. 30, no. 3, 1998.

2. Hall, E. "Bullet Markings from Consecutively rifled Shilen DGA Barrels." AFTE *Journal*, vol. 15, Jan., 1983.

Table 1

Gun	Make	Model	Serial Number	Caliber	Barrel Length	GRC	Bullet Code	Case Code	Field Origin	DOM
1	Ruger	P89DC	303-31032	9mm	4.5”	(6R) 0.079-0.081”/0.100-0.103”	Z56	Q37	Salt Lake City, UT	Jul-91
2	Ruger	P89	312-60197	9mm	4.5”	(6R) 0.078-0.081”/0.099-0.102”	A41	V39	Charlotte, NC	Jun-98
3	Ruger	P89	309-85116	9mm	4.5”	(6R) 0.078-0.081”/0.100-0.103”	B34	X11	Salt Lake City, UT	Dec-94
4	Ruger	P89	305-08156	9mm	4.5”	(6R) 0.079-0.081”/0.101-0.103”	F26	T74	Kansas City, KS	Mar-93
5	Ruger	P89	304-96450	9mm	4.5”	(6R) 0.078-0.080”/0.100-0.103”	C67	L28	Atlanta, GA	Apr-93
6	Ruger	P89	310-67847	9mm	4.5”	(6R) 0.078-0.080”/0.101-0.103”	G91	D73	Salt Lake City, UT	Apr-96
7	Ruger	P89	304-95295	9mm	4.5”	(6R) 0.078-0.080”/0.100-0.103”	E59	W51	Washington, DC	Apr-93
8	Ruger	P89	304-46959	9mm	4.5”	(6R) 0.078-0.080”/0.099-0.102”	X77	H38	Charlotte, NC	Jul-92
9	Ruger	P89DC	303-82472	9mm	4.5”	(6R) 0.078-0.080”/0.100-0.103”	S23	J66	Louisville, KY	Jan-92
10	Ruger	P89	304-96450	9mm	4.5”	(6R) 0.078-0.080”/0.100-0.103”	P87	K97	Atlanta, GA	Apr-93

Table 2

	Q37		V39		X11		T74		L28		D73		W51		H38		J66	
Q37																		
V39	NC:8																	
X11	NC:8		NC:8															
T74	NC:7 NI: 1		NC:7 NI: 1		NC:8													
L28	NC:7 NI: 1		NC:7 NI: 1		NC:7 NI: 1		NC:8											
D73	NC:8		NC:8		NC:8		NC:7 NI: 1		NC:8									
W51	NC:7 NI: 1		NC:7 NI: 1		NC:7 NI: 1		NC:8		NC:8		NC:7 NI: 1							
H38	NC:8		NC:8		NC:8		NC:8		NC:8		NC:8		NC:8					
J66	NC:8		NC:8		NC:8		NC:7 NI: 1		NC:7 NI: 1		NC:8		NC:7 NI: 1		NC:8			
K97	NC:7 NI: 1		NC:7 NI: 1		NC:7 NI: 1		NC:8		I:7 NC:1		NC:7 NI: 1		NC:8		NC:8		NC:7 NI: 1	

Number Participants:8

Number of Comparisons: 360
False Identifications: 0
False Eliminations:0
No Conclusions: 335
True Identifications: 8 / Identifications: 7
True Eliminations (exlcusions): 352/ Non-Identifications: 18

Answers

Q37	NI/NC
V39	NI/NC
X11	NI/NC
T74	NI/NC
L28	I(K97)
D73	NI/NC
W51	NI/NC
H38	NI/NC
J66	NI/NC

Table 3

	Z56		A41		B34		F26		C67		G91		E59		X77		S23
Z56																	
A41	NC:8																
B34	NC:8		NC:8														
F26	NC:8		NC:8		NC:8												
C67	NC:8		NC:8		NC:8		NC:8										
G91	NC:8		NC:8		NC:8		NC:8		NC:8								
E59	NC:8		NC:8		NC:8		NC:8		NC:8		NC:8						
X77	NC:8		NC:8		NC:8		NC:8		NC:8		NC:8		NC:8				
S23	NC:8		NC:8		NC:8		NC:8		NC:8		NC:8		NC:8		NC:8		
P87	NC:8		NC:8		NC:8		NC:8		I:6 NC:2		NC:8		NC:8		NC:8		NC:8

Number of Participants: 8
Number of Comparisons : 360
False Identifications: 0
False Eliminations: 0
No Conclusions: 354
True Identifications: 8 / Identifications: 6
True Eliminations (exclusions): 352 / Non-Identifications: 0

Answers	
A41	NI/NC
B34	NI/NC
F26	NI/NC
C67	I (P87)
G91	NI/NC
E59	NI/NC
X77	NI/NC
S23	NI/NC
Z56	NI/NC