A part-declared blind testing program in firearms examination

W. Kerkhoff[a], R.D. Stoel[a], E.J.A.T. Mattijssen[a,b], C.E.H. Berger[a,c], F.W. Didden[d], J.H. Kerstholt[e,f]

[a] Netherlands Forensic Institute, PO Box 24044, 2490 AA The Hague, The Netherlands
[b] Radboud University Nijmegen, Behavioural Science Institute, PO Box 9104, 6500 HE Nijmegen, The Netherlands
[c] Leiden University, Institute for Criminal Law and Criminology, PO Box 9520, 2300 RA Leiden, The Netherlands
[d] Central Unit, Dutch National Police, PO Box 100, 3970 AC Driebergen, The Netherlands
[e] University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
[f] TNO, PO Box 23, 3769 ZG Soesterberg

Abstract

In 2015 and 2016 the Central Unit of the Dutch National Police created and submitted 21 cartridge case comparison tests as real cases to the Netherlands Forensic Institute (NFI), under supervision of the University of Twente (UT). A total of 53 conclusions were drawn in these 21 tests. For 31 conclusions the underlying ground truth was ‘positive’, in the sense that it addressed a cluster of cartridge cases that was fired from the same firearm. For 22 conclusions the ground truth was ‘negative’, in the sense that the cartridge cases were fired from different firearms. In none of the conclusions, resulting from examinations under casework conditions, misleading evidence was reported. All conclusions supported the hypothesis reflecting the ground truth. This article discusses the design and results of the tests in more detail.

Highlights

- The performance of the Netherlands Forensic Institute’s (NFI) Firearms Section was tested in a part-declared validity study.
- The 21 tests were prepared by external parties and submitted as real forensic cases to the NFI.
- In 18 tests the examiners could not distinguish the tests from real cases.
- All 50 conclusions supported the hypothesis reflecting the ground truths of the 21 tests.

Keywords: part-declared testing; blind testing; fake cases; bullet and cartridge case comparison; proficiency test.

1. Introduction

Modern firearms fire cartridges, each one typically consisting of a projectile (bullet), propellant (powder charge), and igniter (primer) held together by a cartridge case. Most modern firearms are automatic and/or semi-automatic. When a cartridge is fired in such a firearm, the bullet is fired at the target through the barrel and the cartridge case is expelled from the firearm. The firearm typically marks the bullet and the cartridge case with striations or impressions. The distribution, shape and size of these striations and impressions may vary per individual firearm. Comparing these marks with a comparison microscope can give information on the question whether two or more cartridge cases or bullets were fired from the same firearm or whether they were fired from a specific firearm. This type of examination is referred to as e.g. ‘cartridge case and bullet comparison’, ‘forensic firearms examination’, and ‘forensic firearm identification’ in the literature [1]. This discipline is a feature-comparison method from which the validity has been critically highlighted in the 2009 NAS-report [2] and in the 2016 PCAST-report [3].

Periodic ‘blind’ testing of examiners can help to assess the validity of conclusions drawn from cartridge case and bullet comparisons. It also offers the possibility to provide feedback to examiners working under casework conditions. ‘Blind’, ‘double blind’, ‘declared double-blind’ and ‘external blind’ testing has been referred to in various ways in the literature [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. Kerkhoff et al. [4] and Stoel et al. [5] have used the term ‘double blind’ to denote studies in which examiners were not aware that they were being tested. In medical science, the term double blind is well established for clinical trials where both the tested subjects and the personnel administering the test samples have no knowledge of the test design, e.g. in the sense that both do not know which sample e.g. contains the tested drug or a placebo. In line with this definition, the term double blind has been used in forensic literature concerning firearms examination by Smith et al. [6], Stroman [7], and Bunch and Murphy [8] to denote studies in which both the tested firearms examiners and the administrators of the tests had no way of knowing the correct outcome of the tests. Another feature of these three studies was that extra care was taken to make the tests as realistic as possible. To distinguish her study from other studies where the tested examiners were not aware that they were being tested, Stroman [7] labelled her study a ‘declared double blind’ test. In the current study, as in the earlier one by Kerkhoff et al. [4] (then labelled a ‘double-blind’ study) the tested examiners knew they could be tested but did not know whether or not they were working on a test or a real case at the time of examination. To distinguish the current study from the studies by Smith et al. [6], Stroman [7], and Bunch and Murphy [8], and building on the definition used by Stroman, the tests deployed in the current study will be labelled ‘part-declared blind’ tests. This type of test is defined here as a test where the tested examiner does not know the ground truth of the test, knows that he or she can be tested, but does not know whether or when he or she is actually working on a test or on a real case. The police agencies that served as administrators of the tests knew the ground truth of the tests. The ‘ground truth’ is defined here, following the SWGFAST [14] definition, as ‘definite knowledge of the actual source of cartridge cases and bullets’, and is used for instance by Thompson et al. [15] in this sense. Contact between the administrators of the tests and the tested examiners was limited to the examiner receiving a written request to examine the submitted evidence and the administrator receiving a written report in return, as in real cases. The current study was publically announced in advance via a Letter to the Editor in Science & Justice by Stoel et al. [5], then still referring to 'double-blind' tests. With the announcement, the authors wanted to express their commitment to publish their results, regardless of the outcome. This was done in order to prevent the possible future problem of publication bias, that would arise when unfavourable results from the current and similar studies would not be published. In that event, an analysis of published results will be biased because it will only include the more favourable results.

2. Study design and set-up

2.1. Improvements on the earlier study

The current study was built on the experiences from an exploratory study [4], conducted in 2010, 2011, and 2012. In this exploratory study, 10 cartridge case and bullet comparison tests were prepared and submitted by various police agencies as regular cases to the NFI under supervision of the University of Amsterdam (UvA). The results of the tests were evaluated with the VU University Amsterdam (VUA). A total of 29 conclusions were drawn in the 10 tests. For 19 conclusions the ground truth was 'positive', in the sense that the submitted cartridge cases or bullets were either fired from the questioned firearm or from one and the same firearm (in tests where no firearm was submitted). For 10 conclusions the underlying ground truth was 'negative'. In none of the conclusions misleading evidence was reported, in the sense that all conclusions supported the hypothesis reflecting the ground truth. The current study included more tests which resulted in more conclusions. In contrast with the exploratory study, a choice was made to focus on a single case type and a single evidence type. The blinding was improved by involving only one of the NFI's firearms examiners in planning the study, instead of the three that were involved in the exploratory study. The blinding was further improved by sending in tests through an unsuspected source, as will be explained in Section 2.4. Last, a more in depth assessment of the effectivity of the blinding procedure was performed.

2.2. Case type

The case type selected for the current study typically consists of ammunition parts (bullets or cartridge cases) from minor incidents (e.g. vandalism) without victims or suspects. The submitting agencies request to enter the ammunition parts in the open case file, to be able to link the exhibits to other incidents or test fires from firearms. Before entering the ammunition parts into the open case file, a short, indicative examination is performed to establish whether the ammunition parts were fired from one or more firearms. With cartridge cases, the examiner typically selects and compares the most prominent of the marks with the highest evidential strength and only checks whether the other marks are not inexplicably different. The conclusion of the comparison is reported. The examiner notes that "the results indicate that" the ammunition parts were fired from one firearm, if applicable. If ammunition parts from more than one firearm are received, the number of firearms used and the number of ammunition parts per firearm is reported. Being an indicative examination type, a complete assessment and interpretation of the evidence followed by a conclusion in the form of a likelihood ratio, which is the standard at the NFI in other case types, is not carried out.

2.3. Test scope

For creating the tests, 9mm Luger cartridge cases from 39 Glock pistols and one SIG pistol were selected. Firearms of this calibre are at present the most commonly used ones in shooting incidents in the Netherlands. Apart from an indicative statement about the number of firearms that were used, the type of the used firearm(s), and whether these firearms were used in crimes in the Netherlands is also reported. The assessment of the correctness of the latter two statements was left out of scope in the current study. None of the tests were set-up in such a way that a 'hit' in the open case file should be found. Consequently (and 'correctly'[1]) no hits in the open case file were reported.

2.4. Test preparation and routing

The tests were prepared and distributed by a member of the Central Unit of the Dutch National Police. The aforementioned 40 pistols were used to fire 137 cartridges. A wide variety of ammunition brands with different headstamps was used, as this is commonly encountered in casework in the Netherlands. The 137 cartridge cases (the fired bullets were not collected) were distributed over 21 test sets. See Table 1 for an overview of the test sets.

Table 1. Overview test set-up

Test	Pistol	Cartridge cases	
		Headstamps	Number
T1	P1	*AI*	5
T2	P2	*WIN, DAG*	4
T3	P3	*G.F.L.*	3
	P4	*G.F.L.*	5
T4	P5	*S&B*	4
T5	P6	*S&B*	5
	P7	*S&B*	1
T6	P8	*AI*	3
T7	P9	*S&B*	3
	P10	*GECO*	1
	P11	*S&B*	3
T8	P12	*S&B*	1
	P13	*S&B, MEN*	7
T9	P14	*DAG*	5
	P15	*FNB, CBC*	4
T10	P16	*IMI*	3
	P17	*FFV*	2
T11	P18	*S&B*	1
	P19	*S&B, GECO*	3
T12	P20	*DAG*	5
	P21	*WIN, IMI, DAG*	5
T13	P22	*WIN*	3
T14	P21	*SPEER*	4

[1] Strictly speaking, the ground truth of not finding a hit in the open case file with these tests is not known. The Glock pistols that were used for this study were borrowed from a well-guarded naval depot. The chance that any of these pistols were used in a crime is considered to be very low, but it cannot be ruled out completely.

	P24	*SPEER, R-P*	3
	P25	*IMI*	2
T15	P26	*SPEER, R-P*	3
T16	P27	*SPEER, R-P*	5
T17	P28	*WIN, DAG*	4
T18	P29	*WIN, IMI, DAG*	4
	P30	*WIN*	1
T19	P31	*WIN, IMI, DAG*	5
T20	P32	*FC*	6
	P33	*FC*	7
T21	P34	*DAG*	4
	P35	*DAG*	2
	P36	*LAPUA*	1
	P37	*LAPUA*	1
	P38	*LAPUA*	1
	P39	*LAPUA*	1
	P40	*DAG*	7
Total number of cartridge cases			**137**

Pistol P25, used to prepare test T14, was a SIG pistol. All other pistols were Glock pistols. The cartridge cases with DAG and S&B headstamps were of several varieties (various production years and/or lots, with and without lacquer etc.). Several of the cartridges were corroded with salt water and/or by prolonged atmospheric exposure and some cartridge cases were deliberately damaged (e.g. by being driven over with a vehicle) to mimic casework conditions. Notes were kept on the number of cartridge cases, their headstamps and the firearm(s) they were fired from, for all tests. The sets per test were not selected by their marks. In this way a bias towards either ‘hard’ or ‘easy’ comparisons was prevented. After preparation, the test sets were distributed over various police agencies and submitted as real cases to the NFI within a two year time frame. Eight test sets were submitted as though they were submitted from the Caribbean islands of Bonaire and St. Maarten. The Netherlands has ties with these islands through various constitutional structures. Due to the geographical distance and the difference in time zones, contact between the NFI and Caribbean police agencies is less frequent than for agencies located in The Netherlands. The more independent island of St. Maarten submits its cases to the NFI as a paying customer. For the tests sent in as coming from St. Maarten, a mock signed invoice was prepared and submitted in advance, and approved by uninformed NFI employees. Because of the aforementioned procedure we expected that cases from Caribbean islands would not be believed to be tests by the firearms examiners.

2.5. Monitoring the blinding

Apart from the public announcement [5] mentioned in the introduction, the examiners of the NFI’s Firearms Section were verbally notified of the study. They were told that an unknown number of blind tests could be expected from every possible source for an unrevealed period in time. No further information was given. A questionnaire was appended to all cases (tests and real cases) during the course of the study. The questionnaires were filled out by the examiners after completing each case, stating whether or not they believed the case was a test, and if so, why. After completing an examination, the examiner wrote his or her report as usual and sent it to the agency that submitted the case. The reports of the test cases were collected by the same member of the Dutch National Police that prepared the tests.

2.6. Examiners and their roles

The tests were conducted by ten examiners of the NFI's Firearms Section, see Table 2. The letter code of Table 2 of the current study is consistent with Table 1 in the article of the earlier study [4]. Since examiners D and K left the NFI between the two studies, these letter codes are missing. Examiner L joined the NFI between the earlier and the current study. All examiners except technicians H and L were qualified at the start of the program. Technician H was experienced in other case types but was still in training for this specific type during the program. Technician L was qualified from 01-01-2016 (about mid-program).

Table 2. Information about the tested examiners.

Examiner	Sex	Role	Age (year)	Experience (year)
A	Male	Expert	48	24
B	Male	Expert	31	5
C	Male	Expert	47	21
E	Male	Expert	48	25
F	Female	Expert	43	20
G	Female	Technician	43	7
H	Female	Technician	48	16
I	Male	Technician	51	6
J	Female	Expert	43	15
L	Male	Technician	27	1

A minimum of two examiners were involved in each test, at least one of them being a qualified expert. A qualified technician can complete a case independently and submit it to a qualified expert for review, and vice versa. In both instances, the expert signs the report and is responsible for its content. When the case is performed by either a technician in training or an expert in training, the examination is reviewed by two qualified experts or a qualified technician and a qualified expert, the latter signing the report. When either the initial examiner or the reviewer feels the need, a second opinion can be called for and a second examination can be performed blindly (unaware of the previously drawn conclusions) by a qualified technician or expert. Examiner A was involved in planning the current study and was aware of its scope. He attempted not to be involved in the tests by performing and reviewing as little of this specific type of cases as possible during the study. His involvement in the tests was limited to performing a second opinion in test T10, requested by examiner C. When performing the second opinion, examiner A suspected the case to be a test, but did not know this for certain.

3. Test result evaluation

After receiving a number of cartridge cases, an examiner will cluster them based on their marks, creating clusters of cartridge cases that appear to be fired by the same firearm. The examiner will then try to substantiate this first impression by comparing the marks with a comparison microscope. When applicable, he or she will also try to substantiate that the marks in cartridge cases from one cluster are adequately distinct from those in possible other cluster(s). After completing the examination, the examiner draws up his or her conclusions, naming the number of firearms used to fire the submitted cartridge cases and attributing each cartridge case to a cluster, judged to be fired from one firearm.

The choice was made to evaluate the results of the study by the number of conclusions. The option to evaluate by the number of tests was discarded. This latter type of evaluation does not take into account the number of conclusions drawn per test, each with their own potential for providing misleading evidence (conflicting with the ground truth). There are several ways to define ‘a conclusion’, and when this conclusion is ‘positive’ or ‘negative’. When the marks in two cartridge cases are similar, the conclusion is drawn that the results indicate that the cartridge cases were fired from one firearm. In this study, this statement is seen as one ‘positive’ conclusion which has the potential of being a ‘false positive’ when conflicting with the ground truth. When the marks in two cartridge cases are dissimilar, the conclusion is drawn that the results indicate that the cartridge cases were fired from two firearms. In this study, this statement is seen as one ‘negative’ conclusion with the potential of being a ‘false negative’. When either the number of cartridge cases fired per firearm, or the number of firearms used in one test exceeds two, the definition of ‘one conclusion’ becomes ambiguous. The way to quantify the number of conclusions in the more complex tests of this study is explained by describing test T7 below.

3.1. Example

In test T7, seven cartridge cases were submitted from three different pistols. Three cartridge cases originated from pistol P10, one from pistol P11 and three from pistol P12. Figure 1 visualises the two positive conclusions and the three negative conclusions that should be drawn in test T7, as an example.

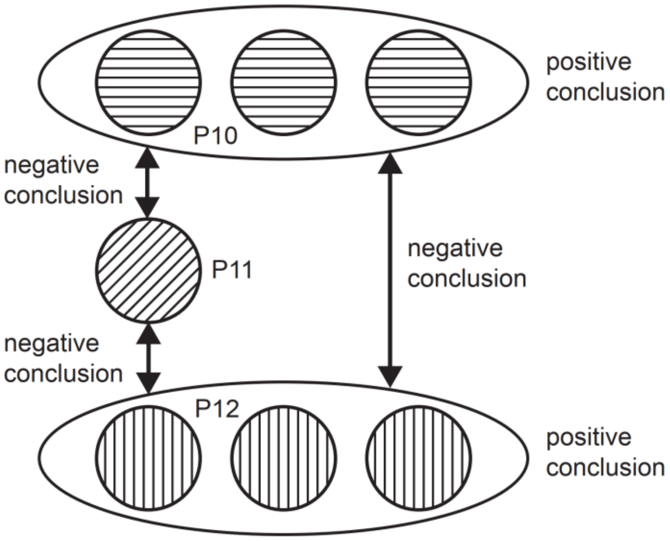

Figure 1. Example of the conclusions that should be drawn, based on the ground truth, in test T7.

A reported conclusion attributing a cluster of cartridge cases to one firearm is counted as one positive conclusion, regardless of the number of cartridge cases the conclusion refers to. In the example, two positive conclusions should be drawn, both with respect to a cluster of three cartridge cases. A reported conclusion attributing (clusters of) cartridge cases to two different firearms is counted as one negative conclusion, regardless of the number of cartridge cases in the respective clusters. In tests where cartridge cases are attributed to three or more firearms, the reported number of firearms is counted as the number of negative conclusions. In the example, three reported negative conclusions should be counted because three different firearms were used to prepare the test.

This way of defining and quantifying conclusions leaves the possibility open that a cartridge case is wrongfully attributed to a cluster, while maintaining the correct number of reported positive and negative conclusions. If, for instance, a cartridge case belonging to cluster P10 would have been wrongfully attributed to cluster P12 in the example, the number of reported positive and negative conclusions would be the same, although the wrongful assignment would constitute both a false positive (attribution to P12) and false negative (failure to assign to P10) with regard to the ground truth of the test. To prevent this problem, an analyses of the results at cartridge case level would be necessary. Analysing the results at this level makes the problem of defining a ‘positive’ and a ‘negative’ conclusion more complex. Furthermore, an analysis at cartridge case level would refer to a much higher number of ‘conclusions’ than the actual number of conclusions stated in the reports that were the results of these tests. For this reason, the aforementioned more concise and conservative form of analysis was chosen. For good measure, the result for each of the 137 cartridge cases was checked and wrongful attributions will be mentioned under Test results.

4. Questionnaire results

Table 3 summarises the questionnaire results from the 21 tests and the 779 real cases that were completed in 2015 and 2016. The questionnaires filled out by examiner A, who was aware of the scope of the study and did not complete tests, were left out of the analysis.

Table 3. Summary of the questionnaire results.

Questionnaire results	21 tests		779 cases	
	Number	%	Number	%
Believed it was a test	3	14.3	8	1.0
Believed it was a case	7	33.3	382	49.0
Didn't know	10	48.6	380	48.8
No/invalid questionnaire	1	4.8	9	1.1

The reasons given in the questionnaires of the 382 real cases, where the examiners believed they had worked on real cases and not tests, are summarised in Table 4. The total number of reasons (416) exceeds the number of questionnaires because more than one reason could be given.

Table 4. Reasons cases were not believed to be tests.

Reason	Number	%
Case involved a victim	90	20.2
From Caribbean island or foreign country	89	20.0
Other examinations requested (e.g. DNA)	58	13.0
Contact with submitting agency or others	42	9.4
Case was a supplement to an earlier case	41	9.2
Known (high-profile) case	38	8.5
Special case type (e.g. reconstruction)	29	6.5
Reference to an older (cold) case	29	6.5
Others	30	6.7

The most frequent reason (five times) given in the eight real cases where the examiners incorrectly believed they had worked on a test, was the case description. Some descriptions apparently lead examiners to believe the case to be faked. Atypical evidence (two), and a lacking court appointment in a murder case were other reasons mentioned. The seven tests T14 to T20 that were believed to be real cases were all received from a Caribbean island, and were not believed to be tests for that reason. In the three tests T5, T8, and T21, that where correctly believed to be tests, the fact that cartridges from different firearms of the same type were received, was mentioned as a reason. The use of atypical ammunition brands was mentioned in two of these tests.

4.1. Discussion questionnaire results

The NFI's Firearms Section removes contextual information from cases prior to examination [16,17]. Written case information is withheld from the examiner in a sealed envelope until after examination and drawing of conclusions. The questionnaires were often filled out after completing the examination, when the contextual information had become known to the examiner. The reasons given in the questionnaires were therefore, at least in part, post-hoc assessments based on information that was not available when examining the cartridge cases. This was the case in five out of the eight real cases that were believed to be tests, where a reference to an odd case description was made. Contextual information does not appear to have played a role in the examiners correct assessment that tests T5, T8 and T21 were indeed tests. Written case information was kept minimal and trivial in these tests and no reference to it was made in the questionnaires. The fact that cartridge cases from more than one firearm were received appeared to have played an important role in correctly identifying tests T5, T8 and T21 as such. Receiving exhibits from two or more different firearms of the same calibre and type is atypical in casework but desirable in a test, because it creates the potential for reporting a false positive. These considerations were apparently taken into account by the tested examiners that were able to identify the three tests.

The key element in making the examiners believe that tests T14 to T20 were real cases, was their disguise as cases from the Caribbean. This particular 'trick' might only work for the Dutch situation, and even there will no longer work after the results of this program are made known to the examiners. Other institutes, contemplating a blind cases program could consider finding similar tricks that will work in their particular situation, or consider not announcing the fact that blind cases are to be expected.

5. Test results

Table 5 gives an overview of the outcomes of the 21 tests, uncorrected and corrected for tests that were correctly believed to be tests. The first examiner in Table 5 was the expert or technician that performed the initial examination. The second examiner was the expert or technician that reviewed the examination. The third examiner either performed a second review (when the initial examination was performed by an expert or technician in training) or performed a blind second examination at the request of the first and/or second examiner. The examiner printed in bold was the expert that signed the report and was responsible for the final conclusion. The examiners that performed tests T5, T8 and T21 correctly believed they worked on a test. The conclusions of those tests were printed in italic and were left out of the 'corrected' totals on the bottom of Table 5. None of the 137 cartridge cases were wrongfully attributed to (a cluster of) cartridge cases fired from a different firearm.

Table 5. Overview of test results

	Examiners			**Conclusions**		
Test	1st	2nd	3rd	Pos.	Neg.	Tot.
T1	G	**B**	-	1	0	1
T2	**J**	E	-	1	0	1
T3	I	**B**	-	2	1	3
T4	I	**J**	-	1	0	1
T5	***C***	*B*	-	*1*	*1*	*2*
T6	L	**C**	I	1	0	1
T7	G	**F**	-	2	3	5
T8	***F***	*E*	-	*1*	*1*	*2*
T9	**C**	G	-	2	1	3
T10	**C**	J	A	2	1	3
T11	**E**	L	-	1	1	2
T12	I	**F**	-	2	1	3
T13	**J**	C	-	1	0	1
T14	H	**F**	J	3	3	6
T15	**J**	E	-	1	0	1
T16	**J**	L	-	1	0	1
T17	**J**	F	-	1	0	1
T18	L	**E**	-	1	1	2
T19	H	**J**	C	1	0	1
T20	H	**J**	C	2	1	3
T21	*L*	***F***	-	*3*	*7*	*10*
Uncorrected totals				**31**	**22**	**53**
Corrected totals				**26**	**13**	**39**

5.1. Discussion of the test results

No misleading evidence was reported, in the sense that all conclusions were consistent with the ground truth in the 21 tests, as given in Table 1. This result does not imply that no misleading evidence is, or has been, reported in practice. The sample size was relatively small (53 uncorrected, 39 corrected conclusions) and a much larger sample would be needed to get a good estimate of the rate of misleading evidence in practice. In order to illustrate the effect of sample size, an 95% confidence interval is computed based on the method of Wilson, following Newcombe [18]. The 0/53 rate of (uncorrected) misleading evidence in this study results in a 95% confidence interval that ranges from zero to 6.8%. This analysis refers to the overall probability on reporting misleading evidence in this specific case type, with cartridge cases as evidence and performed by this specific pool of examiners. Because of the small sample size, a separate analysis on the probability of reporting misleading ‘positive’ or ‘negative’ evidence was not performed. Another problem with a correct assessment of the rate of reporting misleading evidence in real casework is the fact that an assumption must be made on the ratio of the ‘positive’ vs. the ‘negative’ ground truth in real cases.

The discussion above, about an assessment of the overall probability of reporting misleading evidence, was included because in this specific case type no attempt is made to report the strength of the evidence per conclusion, as when reporting a likelihood ratio. The probability of reporting misleading evidence in a conclusion depends, among other things, on the strength of the evidence supporting the respective conclusion. Post-hoc analysis of the tests demonstrated that the strength of the evidence supporting each of the 53 conclusions varied widely in this sample set. Photos 1 and 2 are included to illustrate this point. Photo 1 shows a comparison between two cartridge cases from pistols P16 and P17 from test T10. Photo 2 shows a comparison between two cartridge cases from pistols P20 and P21 from test T12.

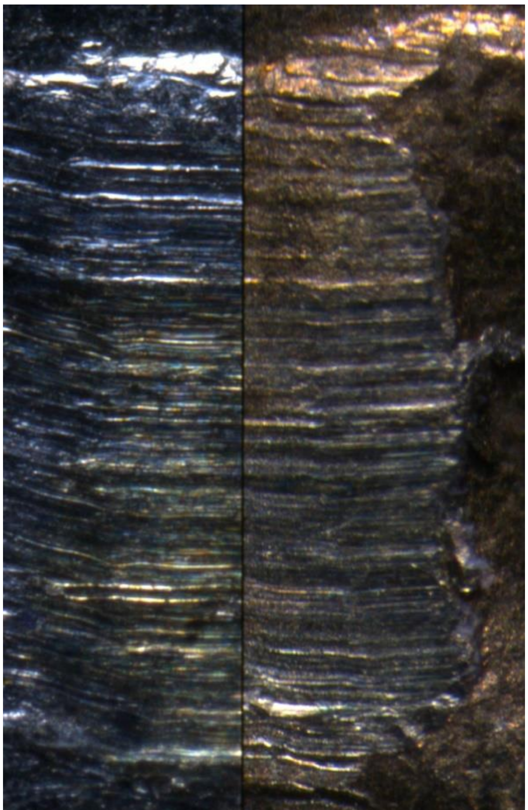

Photo 1. Best alignment of firing pin aperture shear marks in cartridge cases from pistols P16 (left) and P17 (right) from test T10.

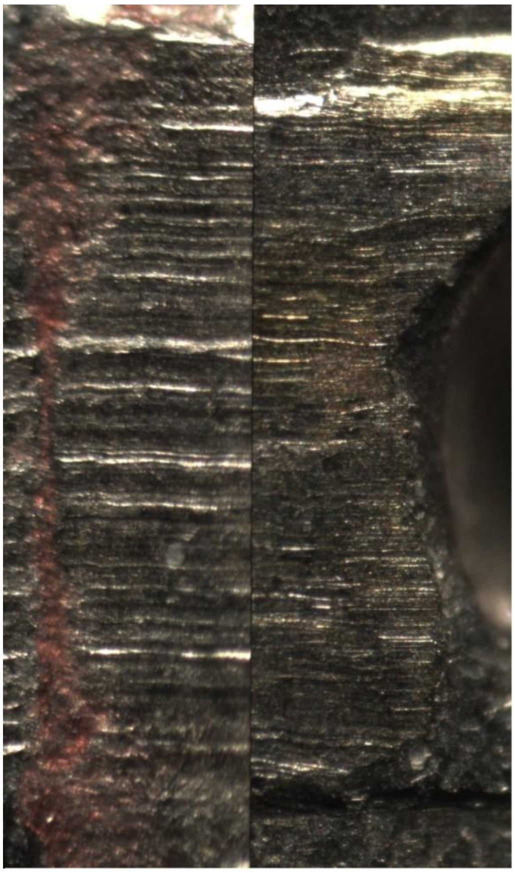

Photo 2. Best alignment of firing pin aperture shear marks in cartridge cases from pistols P20 (left) and P21 (right) from test T12.

On both photos, the striations in the firing pin aperture shear marks are aligned using a comparison microscope. Several of the striations on Photo 1 align, which is not the case on Photo 2. The alignment between the marks in Photo 1 provides some support for the hypothesis that the two cartridge cases where fired from the same pistol (which was not the case). The final conclusion that pointed in the direction that the two cartridge cases were fired from two separate pistols was based on differences in other marks, not visible on Photo 1. But the combined strength of the evidence supporting the negative conclusion in test T10 was lower than in test T12. Therefore, the probability of reporting misleading evidence (in this case a ‘false positive’) was higher in test T10 than in test T12.

Another aspect that might influence the probability of reporting misleading evidence in a test might be the minimum number of pairwise comparisons that are necessary to draw the conclusion(s) in that test. This number increases with the number of cartridge cases and especially with the number of firearms used to prepare a test. The quantification of the minimum number of pairwise comparisons necessary to draw the conclusion(s) in the tests and additional analysis was left out of scope of this study.

6. Overall conclusion and discussion

Proficiency tests and validity studies, consisting of tests prepared by external parties, submitted as real cases and mixed into the regular stream of casework, constitute optimal test conditions for testing both an institute's quality control system and the skills of the tested examiners. The collaborative exercise in the form of the current study might serve as a blue print for such tests. If properly designed and sufficient in numbers, these tests might make a realistic assessment of the rate of misleading evidence in reports possible. One should, however, take into account the limitations implied by sample size and representativeness of the simulated cases in generalizing the results to real case work.

As has been remarked in literature [13], organizing, preparing, and taking such tests is time consuming and costly. Among other things, it involves careful test preparation by external parties that are both disinterested and committed. Yet, as this study demonstrates, it is feasible on a moderate scale. The authors hope that others, who are able to organize such programmes, will do so.

Besides the rate of misleading evidence there is another important aspect of blind testing programs such as the current one. It allows for a form of feedback on performance that cannot be obtained in real case work because the ground truth is always unknown. There is literature focusing on the need for proper feedback and 'deliberate practice' in acquiring and maintaining expert performance.

Experts should be actively and constantly looking for feedback on their performance, preferably the most relevant feedback one can get. As noted by Ericsson et al. [19]: "In the absence of adequate feedback, efficient learning is impossible and improvement only minimal even for highly motivated subjects". The current study showed that it is possible to set up a blind case program that may result in the type of feedback forensic examiners need in order to acquire and maintain their expertise.

7. References

[1] R.S. Bolton-King, Preventing miscarriages of justice: A review of forensic firearm identification, Sci. Justice 56 (2016) 129-142.

[2] Committee on Identifying the Needs of the Forensic Science Community, Committee on Science, Technology and Law Policy and Global Affairs and Committee on Applied and Theoretical Statistics Division on Engineering and Physical Sciences, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington DC, USA (2009).

[3] Executive Office of the President, President's Council of Advisors on Science and Technology (PCAST), Report on Forensic Science in Criminal Courts, September 2016.

[4] W. Kerkhoff, R.D. Stoel, C.E.H. Berger, E.J.A.T. Mattijssen, R. Hermsen, N. Smits, H.J.J. Hardy, Design and results of an exploratory double blind testing program in firearms examination, Sci. Justice 55 (2015) 514-519.

[5] R.D. Stoel, W. Kerkhoff, E.J.A.T. Mattijssen, C.E.H. Berger, Building the research culture in the forensic sciences: Announcement of a double blind testing program, Sci. Justice 56 (2016) 155-156.

[6] T.P. Smith, P.A. Smith, J.B. Snipes, A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework, J. Forensic Sci. 61(4) (2016) 939-946.

[7] A. Stroman, Empirically Determined Frequency of Error in Cartridge Case Examinations Using a Declared Double-Blind Format, AFTE Journal 46-2 (2014) 157-175.

[8] S.G. Bunch, D.P. Murphy, A comprehensive Validity Study for the Forensic Examination of Cartridge Cases, AFTE Journal 35-2 (2003) 201-203.

[9] D.M. Risinger, M.J. Saks, W.C. Thomson, R. Rosenthal, The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion, Calif. Law Rev. (2002) 1-56.

[10] M.J. Saks, D.M. Risinger, R. Rosenthal, W.C. Thompson, Context effects in forensic science, Sci. Justice 44 (2004) 77-90.

[11] M.J. Saks, J.J. Koehler, The Individualization Fallacy in Forensic Science Evidence, Vanderbilt Law Rev. 61 (2008) 199-219.

[12] A. Schwarz, A Systemic Challenge to the Reliability and Admissibility of Firearms and Toolmarks Identification, Columbia Sci. Technol. Law Rev. VI (2005) 1-42.

[13] J.L. Peterson, G. Lin, M. Ho, Y. Chen, R.E. Gaensslen, The Feasibility of External Blind Proficiency Testing. I. Background and Findings, J. Forensic Sci. 48 (2003) 1-9.

[14] SWGFAST (2013). Document #19, Standard Terminology of Friction Ridge Examination (Latent/Tenprint). Retrieved from http://www.swgfast.org/documents/terminology/121124_Standard-Terminology_4.0.pdf on January 13, 2015.

[15] M.B. Thompson, J.M. Tangen, D.J. McCarthy, Human matching Performance of Genuine Crime Scene Latent Fingerprints, Law Hum. Behav. 38(1) (2013) 84-93.

[16] E.J.A.T. Mattijssen, W. Kerkhoff, C.E.H. Berger, I.E. Dror, R.D. Stoel, Implementing context information management in forensic casework: Minimizing contextual bias in firearms examination, Sci. Justice 56-2 (2016) 113-122.

[17] E.J.A.T. Mattijssen, R.D. Stoel, W. Kerkhoff, Minimizing Contextual Bias in Forensic Firearms Examinations, in Wiley Encyclop. of Forensic Sci., eds A. Jamieson and A.A. Moenssens, John Wiley: Chichester. DOI: 10.1002/9780470061589.fsa1117. Published 14th June 2015.

[18] R.G. Newcombe, Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods, Statistics in Medicine 17 (1998) 857-872.

[19] K.A. Ericsson, R.T. Krampe, C. Tesch-Roemer, The role of deliberate practice in the acquisition of expert performance. Psychological Rev. 100 (1993) 363-406.