Design and results of an exploratory double blind testing program in firearms examination

W. Kerkhoff, R.D. Stoel PhD, C.E.H. Berger PhD,
E.J.A.T. Mattijssen MSc, R. Hermsen BSc,
Netherlands Forensic Institute, The Hague, The Netherlands.

N. Smits PhD,
VU University Amsterdam, The Netherlands

H.J.J. Hardy MSc, M. de Jong MSc,
University of Amsterdam, The Netherlands

In 2010, the Netherlands Forensic Institute (NFI) and the University of Amsterdam (UvA) started a series of tests for the NFI's firearms Section. Ten cartridge case and bullet comparison tests were submitted by various external parties as regular cases and mixed in the flow of real cases. The results of the tests were evaluated with the VU University Amsterdam (VUA). A total of twenty-nine conclusions were drawn in the ten tests. For nineteen conclusions 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 ten conclusions the submitted cartridge cases or bullets were either fired from another firearm than the submitted one or from several firearms, in tests where no firearm was submitted. In none of the conclusions misleading evidence was reported, in the sense that all conclusions supported the true hypothesis. This article discusses the design considerations of the program, contains details of the tests, and describes the various ways the test results were and could be analyzed.

Keywords: Blind testing, Fake cases, Bullet and cartridge case comparison, Proficiency test

1 Introduction

The majority of casework in forensic firearms examination consists of cartridge case and bullet comparisons. In this type of examination the marks on bullets and cartridge cases, resulting from the use of firearms, are compared. This is still largely a 'manual', non-automated procedure. The results therefore depend on the skills of the examiners performing the comparisons and on the quality assurance system in which their work is embedded. Periodic 'double blind' testing is an effective tool for quality assurance purposes, and for providing feedback to examiners. In a double blind test, as in other tests, the true origin and connection of the evidence is known. In other words, the 'ground truth' is known to the constructors of the test. The ground truth is defined here, following [1], as definite knowledge of the actual source of marks on cartridge cases or bullets, and is commonly used in the literature (e.g. [2]). In the broader scientific literature the term 'double blind' is used mostly for tests where both the tested persons and those administering the tests do not know the ground truth. In the forensic literature the term double blind testing is more often used for tests where the tested persons do not know the ground truth and are not aware that they are being tested [3, 4]. Saks & Koehler [5] refer to 'closed' testing in this context. Schwarz [6] refers to 'blind' testing. In specialized literature for the field of firearms examination, the term double blind has been used in a study in which both the tested examiners and the administrators were unaware of the ground truth of the tests [7]. The examiners that participated in the aforementioned study were aware that they were being tested, but measures were taken to minimize unwanted effects caused by this awareness. In the current study, the term 'double blind' is used for tests that were mixed in the flow of real cases. The tested examiners could therefore not know when they were being tested, though they could and sometimes did surmise it (as will be described later). The examiners could not know the ground truth of the tests. The double blind testing program ran from January 2010 until January 2013.

2 Objectives of the program

The primary objective of the program was to get an assessment of the error rate in bullet and cartridge case comparison casework. Being able to analyze the cause and nature of errors when they occur was a secondary objective of the program.

2.1 The notion of an error

To be able to discuss the primary objective, the notion of an 'error' needs to be defined, which is not as straightforward as it may seem at first glance. The NFI's firearms section does not make categorical statements on whether or not a bullet or cartridge case was fired from a particular firearm, but reports its conclusion as a likelihood ratio: a verbally expressed assessment of the probability of the findings, under two mutually exclusive hypotheses. The definition of an error is obvious for categorical statements (e.g. 'the bullet was fired from the revolver'), but less so for likelihood ratios. A given likelihood ratio expresses the expert's opinion on the 'evidential value' of the findings. It can be erroneous, in the sense that an examiner overlooked or misinterpreted evidence. But even a correctly assessed evidential value will, by its probabilistic nature, sometimes support a hypothesis that turns out not to be true. The rate at which this occurs is known as the rate of misleading evidence [8]. The rate of misleading evidence decreases as the reported likelihood ratio increases. Misleading evidence ('an error') is defined as: 'reporting support for a hypothesis that is not true'.

2.2 Design considerations

In designing the program, the choice was made to have the tests prepared, distributed and evaluated by one or more disinterested parties. The tests should also be constructed in a way that a bias towards either 'difficult' or 'easy' tests (as compared to the average real case) is avoided. For an assessment of the rate of misleading evidence, when a low rate is expected, difficult tests could be more informative than average or easy tests. However, for an assessment of the overall rate of misleading evidence in casework from a program with only difficult tests, the difficulty of a test should be clearly defined and quantified, and a model would have to be assumed for the relation between the rate of misleading evidence and difficulty of cases and tests. Such a model is not available.

3 The 2010 double blind testing program

3.1 Organizational setting

Five police agencies, that were known to have the required facilities, were requested to produce the tests. Since the NFI has a long-standing working relationship with these agencies they are no truly disinterested parties. Therefore, the University of Amsterdam (UvA) and the VU University Amsterdam (VUA) were approached. The UvA was involved in the design and set-up of the program, coordinated the preparation of the tests, and collected the results. The VUA was involved in evaluating the results.

3.2 Test preparation

The five police agencies were asked to submit bullets and/or cartridge cases with or without firearms and submit them as normal cases. No further instructions were given to select test specimens. The makers were not trained, nor instructed to select potential specimens by their marks. In this way a bias towards either 'difficult' or 'easy' cases was believed to be prevented. The constructors of the tests were asked to include misleading contextual information, but only if they believed they could do so without raising suspicion. If not, they were asked to provide neutral contextual information or none at all. Two UvA scientists assisted in constructing the tests.

3.3 Test routing

After preparation, the tests were submitted as real cases to the NFI. The submitting agencies kept notes about the way the tests were prepared. At the NFI's firearms Section a questionnaire was appended to all cases (tests and real cases) during the course of the program. The questionnaires were filled out by the examiners after completing each case, stating whether or not they believed the case was a test, and 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 police agencies kept the reports, together with their notes about the test construction for future evaluation.

3.4 Examiners

The examiners of the firearms Unit were notified of the program. They were told that an unknown number of blind tests could be expected from every possible source for an unrevealed period. No further information was given. Eleven firearms examiners participated in the program during its three year course. Table 1 lists the age (in years) and years of experience of the participants, at the start of the program (January 1st 2010).

The examiners A, B and C were involved in the design and setup of the program.

3.5 Case types

The blind tests in this program consisted of cases with, and cases without submitted firearms. The questions posed were picked by the police from a list of standardized questions per case type. The questions were:

In cases with a firearm:

1) Comparison: Were the submitted cartridge cases and/or bullets fired from the submitted firearm?
2) Open case file: Was the submitted firearm used in other shooting incidents in the Netherlands?

In cases without a firearm:

1) Comparison: Were the submitted cartridge cases and/or bullets fired from one and the same firearm?
2) Classification: What was the make and model of the firearm(s) that fired these cartridge cases and/or bullets?
3) Open case file: Were the submitted cartridge cases and/or bullets fired from (a) firearm(s) used in other shooting incidents in the Netherlands?

4 Ways to evaluate test results

The primary objective of the program was to get an assessment of the probability of reporting misleading evidence in bullet and cartridge case comparisons. This is the first question in both case types, mentioned above. A report concerning a bullet and cartridge case comparison might contain more than one conclusion. As a rule, one conclusion is drawn per comparison regarding a cluster of similar items, for instance a number of bullets and a firearm of the same caliber. There is also a minimum and a maximum number of comparisons that can be made to draw one conclusion concerning such a cluster. Results from bullet and cartridge case comparisons can be evaluated on several levels.

4.1 Number of tests

A simple way for evaluating the results of a series of tests would be to divide the number of tests where misleading evidence was reported by the total number of submitted tests. This type of evaluation does not take into account the number of submitted items per test and the corresponding number of comparisons that were made. In a test with one submitted firearm and one submitted bullet, there is only one comparison to be made and therefore one chance of reporting misleading evidence. In a test with multiple submitted items the number of comparisons and conclusions is higher, and consequently the probability of reporting misleading evidence is also higher.

4.2 Number of conclusions

Another way to evaluate the results would be to divide the number of misleading conclusions (‘errors’) by the total number of conclusions drawn in a series of tests. Several conclusions can be drawn per test, depending on the submitted items. Each conclusion is a separate statement, having its own potential for containing misleading evidence. Therefore, using the number of conclusions for evaluating results gives a more realistic assessment than using the number of tests.

4.3 Maximum number of possible comparisons

A third way for an evaluation would be to count the number of possible comparisons per test. In a case with one submitted firearm and several submitted cartridge cases the number of comparisons equals the number of questioned cartridge cases. Each one of them is compared against (test fires from) the questioned firearm. This method makes less sense however, when several cartridge cases or bullets are to be compared, to answer the question whether they were fired from the same firearm. If, for example, four bullets are submitted, each of these would in theory be compared to each of the three others. Given that an item cannot be compared against itself and that a comparison of item A vs. item B equals a comparison of item B vs. item A, the maximum number of possible comparisons C_{max} that can be made for n similar items is:

$$C_{max} = (n^2 - n) / 2.$$

In the example of four bullets, the maximum number of comparisons that can be made is six.

4.4 Minimum number of comparisons, necessary to draw the reported conclusions

Evaluating the maximum possible number of comparisons that can be made often overestimates the number of comparisons that are made in practice. In case work, an examiner comparing four bullets often takes one bullet and compares each of the other three bullets to this first one. When the marks in the bullets 1 and 2 are similar and the marks in the bullets 1 and 3 are similar, the marks in the bullets 2 and 3 will also be similar. In that case the examiner reports support for the hypothesis that all bullets were fired from one weapon.

The examiner draws one conclusion concerning four bullets, based on three comparisons. The minimum number of comparisons C_{min} , necessary to reach a conclusion concerning *n* similar items is:

$$C_{min} = n - 1.$$

In the example of four bullets, the minimum number of comparisons that can be made is three.

5 Results

A total of ten double blind tests were submitted and reported. Abstracts of these tests are listed in Appendix A. In three of the tests one or more firearms were submitted together with cartridge cases and/or bullets. In the remaining seven tests only fired cartridge cases and/or bullets were submitted.

5.1 Result comparison

Some tests were constructed in a way that the number of conclusions, and the minimum and maximum number of comparisons, exceeded the number of submitted tests. For example, Test 3 consists of one firearm and one cartridge case, requiring one comparison and subsequently one conclusion. Test 10, on the other hand, consists of eight cartridge cases, requiring a large number of comparisons. In nineteen conclusions the ground truth was ‘positive’, in the sense that the submitted cartridge cases or bullets were either fired from the questioned firearm (in tests where a firearm was submitted) or from the same firearm (in tests where no firearm was submitted). In ten conclusions the ground truth was ‘negative’. See Table 2 for a summary of the results. See Tables 3 and 4 for an overview of the reported conclusions, with a positive and a negative ground truth respectively.

In none of the twenty-nine conclusions drawn in the ten tests misleading evidence was reported, in the sense that it supported the false hypothesis. The reported amount of support for the hypotheses varied. Three conclusions were neutral results, in the sense that no support for either hypothesis was reported.

5.2 Results of classification

In the seven tests where no firearms were submitted, (clusters of) cartridge cases and bullets from eleven firearms were examined. Some classifications were more detailed than others. The most detailed classification mentions the class, caliber, make and one specific model for the firearm that was used. See Table 5 for the results.

In all tests where a class, caliber, make and/or one or several models were mentioned, this was consistent with the ground truth.

5.3 Results open case file

In the ten tests, cartridge cases and/or bullets from twelve firearms were examined against the open case file. The cartridge cases were checked manually and with the Integrated Ballistic Identification System (IBIS) system. The bullets were only checked manually. Three tests were set up in a way that cartridge cases and/or bullets were entered in the open case file and either the respective firearm itself or additional cartridge cases and/or bullets from this firearm were submitted in a later test. In this way, when comparing the evidence of the last case with the open case file, an agreement in the marks with the evidence from this earlier case would ideally be found. finding such agreement is called a ‘hit’ in the open case file. Two out of the three possible hits were made. The missed hit consisted of a submitted Makarov pistol (Appendix, test 5) that was used to fire the three cartridge cases submitted earlier as test 2 (Appendix). The pistols firing pin was deliberately altered between the two tests. The marks from the firing pin and of the breechface are the marks that are used by the IBIS-system and also the most obvious marks to use when manually comparing to the open case file. Breechface marks are notoriously poor on cartridge cases from Makarov pistols. Therefore, the alterations to the firing pin are the best explanation of the fact that the hit was missed. Besides the explanation in this particular case, it is a known fact from case work that potential hits are occasionally missed when checking an open case file. Agreements can be overlooked when marks are poor, when marks have been altered deliberately, or when they have changed by wear over time.

5.4 Results questionnaires

In all, 1118 cases and tests were registered during the program. In eighteen cases no questionnaire could be retrieved, or the questionnaire is incorrect or illegible, to the point that its content could not be used. In the questionnaires, the examiners were asked to indicate whether they believed the case to be a real case or a test. The results of the questionnaires are summarized in Table 6.

The examiner indicated he or she believed the case to be a test in ten questionnaires (two from tests, eight from real cases), mostly based on the contextual information provided with the case. Cases with several cartridge cases and/or bullets with the same caliber but different marks were also, rightly or wrongly, believed to be tests. The reasons that were given in the 619 questionnaires (3 from tests, 616 from real cases) where the examiner indicated he or she believed the case was real and not a test, were more diverse. The most commonly used motivations are listed below, in random order.

- The examiner heard about the case in the media;
- The examiner had been in contact with police personnel or others, concerning the case;
- The case was sent in by a foreign government body (this is done occasionally, on contract);
- Cases containing a large number of exhibits were not believed to be tests, presumably from the assumption that preparing it would be too time-consuming;

- The evidence was contaminated with blood or tissue.

See the Appendix for details of the various reasons mentioned in the questionnaires with the tests.

6 Discussion and outlook

The limited number of tests, conclusions and comparisons in this program limits the possibilities for conclusions on the actual error rate in cartridge case and bullet comparisons. The program does show at the very least, that double blind testing in cartridge case and bullet comparisons is indeed possible.

6.1 The effect of various forms of bias

A special focus for future study could be the influence of bias on cartridge case and bullet comparison. The danger of bias on forensic casework has been emphasized and discussed in a number of studies [e.g. 11–19]. One source of contextual bias is through domain-irrelevant information. The latter typically consists of contextual case information that is irrelevant to the firearms examiner, and is passed along with the evidence. The NFI's firearms Section implemented a context management [20, 21] protocol in 2012. Domain-irrelevant information is filtered out consistently in all cartridge case and bullet comparison cases. The influence of other sources of bias, such as base-rate bias might very well still influence these cases and might be a focus for further study.

6.2 Future (statistical) analyses

Adding results from other programs, set up at the NFI or, hopefully, at other institutes, will increase the number of results and the potential for more advanced and more informative analyses. Only simple aggregated evaluations were used for this study. More advanced modeling is possible for a larger number of results, from either larger scale programs or from meta-analyses of several programs. Such results could be evaluated using a single model. Since the smallest possible unit is a dichotomous (correct/incorrect) outcome, a logistic regression model [9] could be used. Such a model allows for an evaluation per examiner and can include characteristics of each examiner such as the experience, training etc. Because the results are partly interdependent, a multilevel structure could be added to the logistic regression model. Such an enriched model would also allow adding case characteristics such as caliber, firearms type, number of pieces per test etc. One could then test if these characteristics influence performance. In a testing program ran at several institutes it is possible to administer tests to several examiners. These data would allow for, so-called, Item Response Models [10]. These models, that are extensively used in the construction of ability tests in general, allow for a comparison of (a) examiners amongst each other, and (b) cases amongst themselves, and thereby giving valuable information on the performance of the examiners, as well as on the quality of the tests.

Appendix A

Abstracts of the ten tests

Note: in tests 1, 2, 3, 4, 5, 6, 8 and 10 no hit in the open case file was reported. In test 5 the potential of finding a hit was created, by the previous submission of cartridge cases in test 2. In the other seven tests where no hits in the open case file were reported, no potential of finding a hit was created.

Test 1 (case no. 2010.07.08.033, received July 7, 2010, reported July 20, 2010)

Three 9 mm Luger cartridge cases and two bullets were fired from one semi-automatic HS 95 pistol. The cartridge cases and bullets were submitted. Posed questions: were the bullets and cartridge cases fired from one firearm, what was the make and model of this firearm and was this firearm used in crimes in the Netherlands.

Reported results

Very strong support for the hypothesis that the three cartridge cases were fired from one firearm. Strong support for the hypothesis that the two bullets were fired from one firearm. The class, caliber and make were correctly mentioned. No model was mentioned. No hit in the open case file was reported. The examiner (H in Table 1) indicated in the questionnaire he did not think this case was a test.

Test 2 (case no. 2010.09.02.010, received September 1, 2010, reported September 20, 2010)

Three .32 Auto cartridge cases were fired from one semi-automatic Makarov pistol. The cartridge cases were submitted. Posed questions: were the cartridge cases fired from one firearm, what was the make and model of this firearm and was this firearm used in crimes in the Netherlands.

Reported results

Strong support for the hypothesis that the three cartridge cases were fired from one firearm. The class and caliber were correctly mentioned. No make and model could be established. No hit in the open case file was reported. The examiner (E in Table 1) failed to fill out the questionnaire, or the questionnaire was lost.

Test 3 (case no. 2010.09.21.033, received September 15, 2010, reported October 14, 2010)

A Reck, model PK 8008 blank firing pistol was used to fire an 8 mm blank cartridge case. Both the firearm and the cartridge case were submitted. Posed questions: was the cartridge case fired from this firearm and was this firearm used in crimes in the Netherlands.

Reported results

An equal amount of support for the hypothesis that this cartridge case was fired from the submitted firearm, as for the hypothesis that this cartridge case was fired from another firearm. No hit in the open case file was reported. The examiner (D in Table 1) indicated in the questionnaire he did not think this case was a test, because he saw no reason why it should be.

Test 4 (case no. 2010.09.21.015, received September 15, 2010, reported October 20, 2010)

Two .32 Auto cartridge cases, make Geco, were fired from a FN model 115 pistol. A third .32 Auto cartridge case, make S&B, was fired from another FN model 115 pistol. The three cartridge cases were submitted. Posed questions: Were the cartridge cases fired from one firearm, what was the make and model of this firearm and was this firearm used in crimes in the Netherlands.

Reported results

Very strong support for the hypothesis that the two Geco cartridge cases were fired from one firearm. For the S&B cartridge cases, the reported amount of support was about equal for the hypothesis that this cartridge case was fired from the same firearm, as for the hypothesis that this cartridge cases was fired from another firearm. The class, caliber and make were correctly mentioned. The models 1910 and 1922, and models that were derived from these models were mentioned in the report. These include the model 115. No hit in the open case file was reported. The examiner (I in Table 1) indicated in the questionnaire that this case might have been a test. He indicated believing it to be suspicious because of the fact that three cartridge cases, from what might have been two firearms with the same make and model, were received.

Test 5 (case no. 2010.11.19.080, received November 17, 2010, reported December 23, 2010)

The Makarov pistol, that was used to fire the three cartridge cases submitted in Test 2, was altered. The altered gun was then used to fire two more cartridge cases. The firearm and these two cartridge cases were submitted. The question was posed whether the two cartridge cases were fired by the pistol and whether the pistol was used in other crimes in the Netherlands.

Reported results

Very strong support for the hypothesis that the firearm was used to fire the two cartridge cases. No hit in the open case file, with the three cartridge cases from Test 2 or other cases, was reported. The missing of this hit was discussed in the article. The examiner (E in Table 1) indicated in the questionnaire he did not know whether this case was a test or not.

Test 6 (case no. 2010.12.13.041, received December 10, 2010, reported December 30, 2010)

Three cartridge cases were fired from an FN model 1922 pistol. One cartridge case was fired from an FN model 115 pistol. The four cartridge cases were submitted. Posed questions: were the cartridge cases fired from one firearm, what was the make and model of this firearm and was this firearm used in crimes in the Netherlands.

Reported results

Very strong support was reported for the hypothesis that three of the cartridge cases were fired from one firearm. Very strong support was reported for the hypothesis that the fourth cartridge case was fired from another firearm. The class, caliber and make were correctly mentioned. The models 1910 and 1922 and models that were derived from these models were mentioned in the report. These derived models include the model 115. No hit in the open case file was reported. The examiner (C in Table 1) indicated in the questionnaire he suspected that this case was a test. The reasons that were given were the fact that the cartridge cases were remarkably clean, the fact that cartridge casesfrom two different firearms of the same make and caliber were submitted and that the case background information was vague.

Test 7 (case no. 2011.04.26.046, received April 22, 2011, reported May 4, 2011)

One cartridge case and one bullet were fired from the HS 95 pistol that was used to fire the cartridge cases and bullets, submitted in Test 1. Another cartridge case and another bullet were fired from another HS 95 pistol. This latter firearm had false Smith & Wesson markings, which is common with this type of firearm in casework in the Netherlands. The two cartridge cases and bullets were submitted. Posed questions: were the bullets and cartridge cases fired from one firearm, what was the make and model of this firearm and was this firearm used in crimes in the Netherlands.

Reported results

Very strong support was reported for the hypothesis that the two cartridge cases were fired from two different firearms. An equal amount of support was reported for the hypothesis that the two bullets were fired from the same firearm as for the hypothesis that the two bullets were fired from two different firearms. Very strong support was reported for the hypothesis that one of the cartridge cases was fired from the same firearm as the cartridge cases that were submitted in Test 1. Very strong support was reported for the hypothesis that one of the bullets was fired from the same firearm as the bullets that were submitted in Test 1. The class, make and specific model were correctly mentioned for both firearms, including the fact that these firearms can bear false Smith & Wesson markings. No hit in the open case file was reported, other than the reference made to Test 1. The examiner (A in Table 1) indicated in the questionnaire he did not know whether this case was a test or not.

Test 8 (case no. 2011.06.27.106, received June 24, 2011, reported July 11, 2011)

Two cartridge cases were fired from a .25 Auto caliber CZ 45 pistol. The two cartridge cases were submitted. Posed questions: were the cartridge cases fired from one firearm, what was the make and model of this firearm and was this firearm used in crimes in the Netherlands.

Reported results

Very strong support was reported for the hypothesis that the two cartridge cases were fired from one firearm. The class and caliber were correctly mentioned. No make and model could be established. No hit in the open case file was reported. The examiner (G in Table 1) indicated in the questionnaire she did not know whether this case was a test or not.

Test 9 (case no. 2912.09.17.005, received September 14, 2012, reported October 30, 2012)

The HS 95 pistol with false Smith & Wesson markings that was used to fire one of the two cartridge cases and one of the two bullets submitted in Test 7 was used again to fire a cartridge case and a bullet. The other HS 95 pistol, that was used to fire other cartridge cases and the other bullet submitted in Test 7, being the same firearm that was used to fire the three cartridge cases and the two bullets from Test 1, was used to cycle a cartridge. The two firearms, cartridge case, bullet and cartridge were submitted. Posed questions: were the cartridge case and the bullet fired from one of the submitted firearms, was the cartridge cycled in one of the two firearms and were the two firearms used in crimes in the Netherlands.

Reported results

For practical purposes the HS 95 pistol with false Smith & Wesson markings will be designated as "pistol S". The HS 95 pistol without Smith & Wesson markings will be designated as "pistol H". Very strong support for the hypothesis that pistol S was used to fire the submitted cartridge case, and not pistol H. Strong support for the hypothesis that pistol S was used to fire the submitted bullet and not pistol H. Strong support for the hypothesis that pistol H was used to cycle the submitted cartridge and not pistol S. Very strong support for the hypothesis that pistol H was used to fire the cartridge case from Test 1 and one of the cartridge cases from Test 7. Strong support for the hypothesis that pistol H was used to fire the two bullets from Test 1 and the bullet from Test 7. Very strong support for the hypothesis that pistol S was used to fire the other cartridge case from Test 7. No other hit in the open case file was reported. Note: the bullet from pistol S, that was submitted in Test 7, was not entered in the open case file because it had insufficient markings (due to damage to its surface) and was therefore not available for comparison with Test 9. The examiner (I in Table 1) indicated in the questionnaire he did not believe this case to be a test, because of the connections with the other cases.

Test 10 (case no. 2012.12.14.130, received December 12, 2013, reported January 11, 2013)

A Glock 17 pistol was used to fire two cartridge cases. A Glock 26 pistol was used to fire three cartridge cases. Another Glock 26 pistol was used to fire another three cartridge cases. The eight cartridge cases were submitted. Posed questions: were the cartridge cases fired from one firearm, what was the make and model of this firearm and was this firearm used in crimes in the Netherlands.

Reported results

Very strong support for the hypothesis that three different weapons were used to fire the eight cartridge cases. Very strong support for the hypothesis that two of the eight cartridge cases (correct item numbers were mentioned) were fired from the same firearm. Very strong support for the hypothesis that three of the eight cartridge cases (correct item numbers were mentioned) were fired from the same firearm. Very strong support for the hypothesis that the remaining three cartridge cases (correct item numbers were mentioned) were fired from the same firearm. The class, caliber and make were correctly mentioned. No model(s) was(were) mentioned. No hit in the open case file was reported. The examiner (G in Table 1) indicated in the questionnaire she was convinced she was working a real case during the examination, but had second thoughts when she read the context information after the examination.

References

[1] SWGFAST, Document #19, Standard Terminology of Friction Ridge Examination (Latent/Tenprint) Retrieved from https://www.nist.gov/document/swgfaststandard-terminology40121124pdf (on January 13, 2015).

[2] M.B. Thompson, J.M. Tangen, D.J. McCarthy, Human matching performance of genuine crime scene latent fingerprints, Law Hum. Behav. 38 (2013) 84–93.

[3] 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. 90 (2002) 1–56.

[4] M.J. Saks, D.M. Risinger, R. Rosenthal, W.C. Thompson, Context effects in forensic science, Sci. Justice 43 (2003) 77–90.

[5] M.J. Saks, J.J. Koehler, The individualization fallacy in forensic science evidence, Vanderbilt Law Rev. 61 (2008).

[6] A. Schwarz, A systemic challenge to the reliability and admissibility of firearms and toolmarks identification, Columbia Sci. Technol. Law Rev. VI (2005).

[7] S.G. Bunch, D.P. Murphy, A comprehensive validity study for the forensic examination of cartridge cases, AFTE J. 35 (2003).

[8] Richard M. Royall, Statistical Evidence: A Likelihood Paradigm, Chapman & Hall, London, 1997.

[9] A.A. Agresti, Categorical data analysis, 3rd edition John Wiley & Sons, 2013.

[10] S.E. Embertson, S.P. Reise, Item response theory for psychologists, Mahwah N.J., Lawrench Erlbaum Associates, 2000.

[11] I.E. Dror, D. Charlton, A.E. Péron, Contextual information renders experts vulnerable to making erroneous identification, Forensic Sci. Int. 156 (2006).

[12] O.o.t.I.G. U.S. Department of Justice (Ed.), A Review of the FBI's Handling of the Brandon Mayfield Case, 2006.

[13] J. Kerstholt, A. Eikelboom, T. Dijkman, R. Stoel, R. Hermsen, B. van Leuven, Does suggestive information cause a confirmation bias in bullet comparisons? Forensic Sci. Int. 198 (2010).

[14] J. Kerstholt, R. Paashuis, M. Sjerps, Shoe print examinations: effects of expectation, complexity and experience, Forensic Sci. Int. 165 (2007).

[15] B. Schiffer, Ch. Champod, The potential (negative) influence of observational biases at the analysis stage of fingermark individualization, Forensic Sci. Int. 167 (2007).

[16] M.J. Saks, D.M. Risinger, R. Rosenthal, W.C. Thompson, Context effects in forensic science; a review and application of the science of science to crime laboratory practice in the United States, Sci. Justice 43 (2003) 77–90.

[17] M.J. Saks, J.J. Koehler, The coming paradigm shift in forensic identification science, Science 309 (2005).

[18] M.J. Saks, H. VanderHaar, On the "general acceptance" of handwriting identification principles, J. Forensic Sci. 50 (2005).

[19] J.J. Koehler, fingerprint error rates and proficiency tests: what they are and why they matter, Hastings Law J. 59 (2008) 1077–1100.

[20] R.D. Stoel, C.E.H. Berger, W. Kerkhoff, E.J.A.T. Mattijssen, I.E. Dror, Minimizing contextual bias in forensic casework, in: K.J. Strom, M. Hickman (Eds.), Forensic Science and the Administration of Justice, SAGE Publications 2013, pp. 67–86.

[21] E.J.A.T. Mattijssen, R.D. Stoel, C.E.H. Berger, W. Kerkhoff, Minimizing contextual bias in forensic firearms examinations, in: A. Jamieson, A. Moenssens (Eds.), Wiley Encyclopedia of Forensic Science, Wiley, 2015 In Press. 519 W. Kerkhoff et al. / Science and Justice 55 (2015) 514–519.

Tables

Table 1

Participating examiners

Examiner	Age	Exp.
A	43	19
B	26	0
C	42	17
D	29	5
E	43	20
F	38	15
G	38	2
H	43	11
I	46	1
J	38	10
K	61	38

Table 2

Summary

Number of tests	10
Number of conclusions regarding the different (clusters of) submitted items	29
Minimum number of comparisons, necessary to draw the 29 reported conclusions	31
Maximum number of comparisons, that could be made in the 10 tests	66

Table 3

Results for the 19 conclusions with a positive ground truth

Reported very strong support for positive hypothesis	14
Reported strong support for positive hypothesis	4
Reported equal amount of support for positive hypothesis and negative hypothesis	1
Reported support for negative hypothesis (misleading evidence)	0

Table 4

Results for the ten conclusions with a negative ground truth

Reported very strong support for negative hypothesis	6
Reported strong support for negative hypothesis	2
Reported equal amount of support for negative hypothesis and positive hypothesis	2
Reported support for positive hypothesis (misleading evidence)	0

Table 5

Classification results

Only class of weapon (for instance ‘semi-automatic pistol’) and caliber were mentioned	2
Class, caliber and make were mentioned	4
Class, caliber, make and several models were mentioned	4
Class, caliber, make and one specific model was mentioned	1

Table 6

Questionnaire results

	Blind tests		Real cases	
	Nr.	%	Nr.	%
Believed it was a test	2	20	8	0.7
Believed it was a case	3	30	616	55.6
Didn’t know	4	40	467	42.2
No/invalid questionnaire	1	10	17	1.5
Total	10	100	1108	100