Science and Justice 55 (2015) 514–519

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

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

W. Kerkhoff [a,*], R.D. Stoel [a], C.E.H. Berger [a,d], E.J.A.T. Mattijssen [a], R. Hermsen [a], N. Smits [b], H.J.J. Hardy [c]

[a] Netherlands Forensic Institute, Netherlands Forensic Institute, NFI, The Hague, The Netherlands
[b] VU University Amsterdam, VUA, Amsterdam, The Netherlands
[c] University of Amsterdam, UvA, Amsterdam, The Netherlands
[d] Leiden University, Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 23 November 2012
Received in revised form 8 June 2015
Accepted 12 June 2015

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

ABSTRACT

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.

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

* Corresponding author at: P.O. Box 24044, 2490 AA, The Hague, The Netherlands.
E-mail address: w.kerkhoff@nfi.minvenj.nl (W. Kerkhoff).

http://dx.doi.org/10.1016/j.scijus.2015.06.007
1355-0306/

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.

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