Validity and reliability of forensic firearm examiners

Erwin J.A.T. Mattijssen[a,b], Cilia L.M. Witteman[a], Charles E.H. Berger[b,c], Nicolaas W. Brand[b], Reinoud D. Stoel[b]

[a] Radboud University Nijmegen, Behavioural Science Institute, PO Box 9104, 6500 HE Nijmegen, The Netherlands

[b] Netherlands Forensic Institute, PO Box 24044, 2490 AA The Hague, The Netherlands

[c] Leiden University, Institute for Criminal Law and Criminology, PO Box 9520, 2300 RA Leiden, The Netherlands

Abstract

Forensic firearm examiners compare the features in cartridge cases to provide a judgment addressing the question about their source: do they originate from one and the same from two different firearms? In this article, the validity and reliability of these judgments is studied and compared to the outcomes of a computer-based method. The features we looked at were the striation patterns of the firing pin aperture shear marks of four hundred test shots from two hundred Glock pistols, which were compared by a computer-based method. Sixty of the resulting 79,800 comparisons were shown to 77 firearm examiners. They were asked to judge whether the cartridge case had the same source or a different source, and to indicate the degree of support the evidence provided for those judgments.

The results show that the true positive rates (sensitivity) and the true negative rates (specificity) of firearm examiners are quite high. The examiners seem to be slightly less proficient at identifying same-source comparisons correctly, while they outperform the used computer-based method at identifying different-source comparisons.

The judged degrees of support by examiners who report likelihood ratios are not well-calibrated. The examiners are overconfident, giving judgments of evidential strength that are too high. The judgments of the examiners and the outcomes of the computer-based method are only moderately correlated.

We suggest to implement performance feedback to reduce overconfidence, to improve the calibration of degree of support judgments, and to study the possibility of combining the judgments of examiners and the outcomes of computer-based method to increase the overall validity.

1 Introduction

Judicial systems rely on the forensic science disciplines to provide scientific evidence that can be used in the court of law [1]. One of these disciplines is forensic firearm examination. The main role of firearm examiners is to provide evidence about the source of cartridge cases and bullets that are recovered after a shooting incident. These cartridge cases and bullets contain marks with features – striations and impressions – that originate from components of the firearm with which they were fired. Those features can be compared with the features in reference shots fired with a submitted firearm. The results of such a comparison are used to provide a judgment about the question whether the shots were fired with the submitted firearm or with a different firearm. When there is no submitted firearm, the features of different cartridge cases from the crime scene can also be compared to judge whether those were fired with the same firearm or with different firearms.

Comparing features is traditionally done by examiners, acting as the main instrument of analysis and interpretation [2-4]. Even though courts often treat the testimonies of examiners as impartial [3, 5], they are often criticized for their lack of scientific rigor [1, 6-8]. In particular, forensic disciplines that rely heavily on feature comparison, such as firearm examination, would benefit from the development of a research culture where the goal is to develop a well-established scientific foundation, and where judgments are substantiated by empirical research in addition to training and experience [8, 9]. Such research should focus on the validity and reliability of methods and their application [8]. Several avenues of research have been proposed. These include the development of more objective computer-based methods [8]; the quantification of the variability of features coming from the same or from different sources [1]; the shift from the false idea that judgments could be based on uniqueness of features to establishing their evidential strength and to report judgments in probabilistic terms [8]; the determination of the validity of judgments [1, 8], preferably by double-blind proficiency tests in casework [9, 10]; the implementation of context information management to minimize the risks of cognitive bias [1, 3, 6, 8, 9, 11-14], including e.g., (linear) sequential unmasking, where the evidential material is not examined simultaneously with the reference material, but before examination of and comparison with the reference material [3, 15, 16]; the management of case information [17-19] and blind peer review [3, 6, 7, 20-23].

In firearm examination most scientific effort has been on the determination of the validity of judgments and on the development of more objective computer-based methods. Multiple studies have been set up with the aim to show that examiners are able to correctly judge whether a cartridge case or bullet was fired with a specific firearm or not [e.g., 24, 25-31]. Overall, these studies report low error rates when comparing the judgments of examiners with the ground truth, which is the known correct answer (same-source or different-source judgment) based on the study design. Although these results seem promising, it is unsure how these results of experiments relate to the validity that can be expected during actual casework. There are several limitations of these studies, such as dependencies between judgments due to study designs, the use of closed sets with reference specimens for each questioned sample, and the relevance of the specimens used when compared to normal casework. Blind proficiency tests performed in the normal case flow seem to overcome these issues [10, 32].

More objective computer-based methods have been developed [e.g., 33, 34-47]. These studies usually rely on 3D surface topography measurements of the striation or impression patterns in fired cartridge cases or bullets. The measurements are then compared to each other by computer algorithms, resulting in a comparison score that gives some degree of similarity. Depending on the applied interpretation paradigm these scores are then used directly to decide (implicitly assuming some prior odds and cost/benefit of wrong/right decisions) about the source of the questioned cartridge cases or bullets and to assess an error rate [e.g., 36, 38, 39, 42], or to determine the evidential strength [e.g., 33, 48-50]. As a measure of the evidential strength, the likelihood of the comparison score is assessed for mutually exclusive propositions, for example *H1*: the two cartridge cases were fired with the same firearm, and *H2*: the two cartridge cases were fired with different firearms. The ratio of these likelihoods provides the evidential strength, the likelihood ratio (LR) [51]. An LR above 1 represents support for *H1* over *H2*, and a LR below 1 means support for *H2* over *H1*. These two interpretation paradigms correspond to two currently applied reporting formats for examiner judgments. In one of these, categorical same-source decisions are made when the features are in “sufficient agreement” [52], according to the scientifically-flawed individualization principle, and in the other the likelihoods of the features are assessed given two propositions resulting in an opinion on the strength of the evidence [53-56].

Independent of the applied interpretation paradigm, both the examiners and the computer-based methods take into account the degree of similarity of the features in e.g., two cartridge cases when providing information about their source. Ceteris paribus, a higher degree of similarity will provide stronger support for the proposition that cartridge cases are from the same source. Because the examiners and the computer-based method consider similar features and apply (subjective) comparison algorithms based on the same metric of degree of similarity, it seems reasonable to expect that the outcomes of the two are coherent, in the sense that the judged degrees of support of examiners are positively correlated to the comparison scores from the computer-based method. Neither the examiner judgments nor the outcomes of a computer-based method can be considered as a golden standard. The validity of examiner judgments needs to be further determined [1, 8], while the computer-based methods are still in an experimental stage.

The aims of this study are to assess the validity and reliability of source judgments by examiners and the validity of a computer-based method, to determine the relation between examiners' judgments and the outcomes of a computer-based method, and to determine how calibrated the judged degrees of support of firearm examiners are. To do this we focus on one of the marks that is present in cartridge cases fired with Glock pistols, the firing pin aperture shear mark. The features of this mark are striations on the primer cup of the cartridge case that are caused by the margins of the firing pin aperture of the breechface when the barrel unlocks from the slide.

2 Materials and methods

In this section we first provide information about the firearms and ammunition we used for this study. Secondly, we discuss the computer-based method we used, where we provide information about the data-acquisition, data pre-processing, striation pattern comparison, and the calculation of likelihood ratios. Then we discuss the acquisition of examiner judgments, where we provide details about the subjects and the study design. We conclude this section with an overview of the analyses that we will perform.

2.1 Firearms and ammunition

We used a total of 200 9mm Luger Glock pistols. These firearms were seized in the Netherlands. We chose Glock pistols because they are prevalent in shooting incidents across the world, ensuring that participating firearm examiners are familiar with their features. We fired two shots with each of the firearms, using Fiocchi 9mm Luger ammunition with nickel colored primer cups.

2.2 Computer-based method

2.2.1 Data acquisition

We acquired two- and three-dimensional measurements of the striations of the firing pin aperture shear marks of the four hundred cartridge cases. For the 2D measurements we took digital images using a Leica FS C microscope [57] combined with a Leica DFC490 [58] digital camera. The magnification was set at 60×, resulting in exported images of 3264x2448 pixels, with an RGB color depth of 8 bits per channel without compression. The first author (a certified firearm examiner) visualized the striation patterns using oblique lighting, optimized to show as many of the striations as possible while avoiding overexposure. For the 3D measurements the firing pin aperture shear marks were first cast using gray Forensic Sil [59]. This was done to reduce measurement noise caused by light reflecting from the metal primer cup. We acquired the 3D surface topographies using the Alicona InfiniteFocusSL [60], which uses white light focus variation. The acquisition parameters were set to 2 µm lateral resolution and 200 nm vertical resolution, using a 20× magnification objective, and a full 360° ring light.

2.2.2 Data pre-processing and the comparison of striation patterns

We manually cropped the exported 2D and 3D data files to select the striation patterns in the firing pin aperture shear mark to be considered for further comparison. This was necessary because the data files contained more information than just the striations patterns of interest. By selecting only the striation patterns to be considered for the comparison we ensured that the computer-based method and the examiners would not be privy to information outside the scope of this study (e.g., part of the breechface and firing pin impressions as can be seen the 2D and 3D measurements in Figure 1). We paid special attention to select exactly the same parts of the striation patterns for the 2D and 3D data. Because of differences in acquisition resolution, this resulted in a length of the cropped area of 120 pixels for the 2D data and 160 pixels for the 3D data. The width of the cropped area was determined by the bounds of the firing pin aperture shear marks. This resulted in cropped areas of approximately 0.1 by 1 mm. To be able to compare the striation patterns, the acquired 2D and 3D data need to be pre-processed into striation profiles. Such a profile is an averaged 1D profile from which the overall shape and noise outside the set band-passes of 250 µm and 5 µm, respectively, are removed. The 1D profile represents the striation pattern along the direction of its creation (length of the cropped area). After cropping, the data were pre-processed into a 1D profile following an automated approach, using scripts developed in-house [61].

Before determining the similarity, a multi-scale registration framework including two degrees of freedom: translation and scaling, was used to align two striations patterns. Translation was set to a maximum of 0.1 mm and is used to move the profiles relative to each other. Scaling was set to a maximum of 3.0 % and is used to correct for compression between acquired profiles by allowing for stretching and shrinking. After alignment, the degree of similarity between two profiles is determined by a comparison score, the cross-correlation. This comparison score ranges from -1 to 1, with -1 representing maximum negative correlation, 1 representing maximum positive correlation, and 0 representing no correlation. Please refer to Baiker et al. 2014 [61] for additional details regarding the applied pre-processing steps, profile alignment, and comparison which were originally developed for the comparison of toolmark striation patterns [61-64]. Figure 1 shows a schematic representation of the steps taken to compare two firing pin aperture shear marks from the same source (one firearm) using both 2D and 3D measurements.

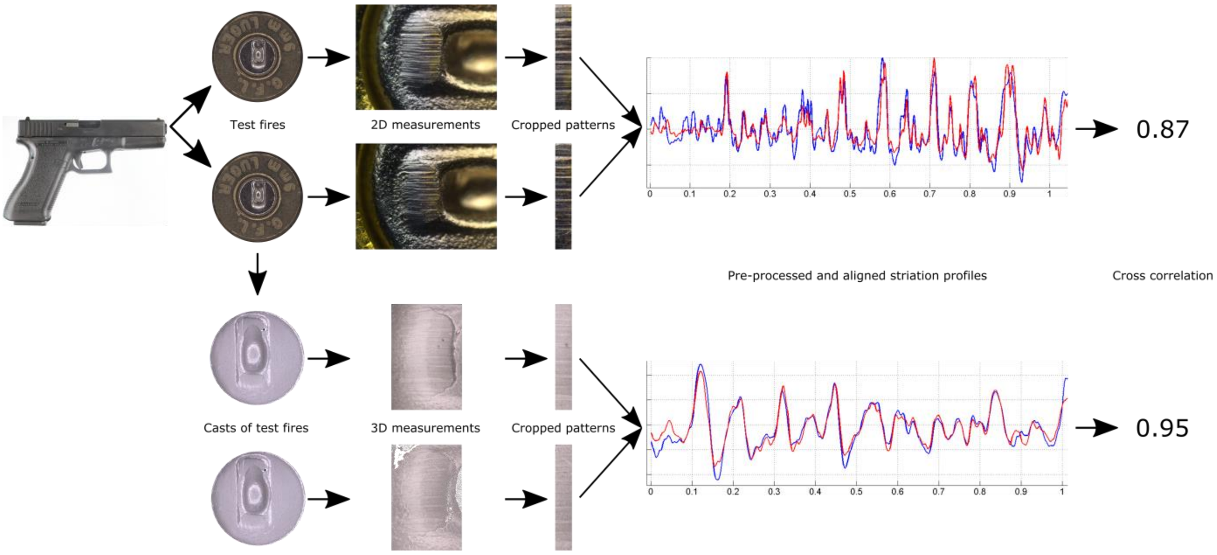

Figure 1

Schematic representation of the steps taken to compare two firing pin aperture shear marks from the same source using both 2D and 3D measurements. From left to right: tests fires are created and cast, 2D measurement from the test shots and 3D measurements from the casts are acquired, the striation patterns to compare are selected, the data is pre-processed and the resulting 1D striation profiles are aligned, the comparison score is calculated.

2.2.3 LR calculation

All acquired striation profiles were compared to each other. This resulted in 200 same-source and 79600 ((400×399/2) - 200) different-source comparison scores for both the 2D and 3D data. The distributions of the same-source and different-source comparison scores are modeled by kernel density estimation, using the built-in density function in R [65] with bandwidths for same-source and different-source distributions set at 1.5. The resulting distributions are generic for cases from the relevant population and can be used for common-source questions such as: are these two cartridge cases fired with the same firearm? Here, the reproducibility of a specific firearm in a case is not considered as it is in specific-source questions such as: is the cartridge case fired with the submitted firearm? To calculate a similarity-only score-based LR for a specific comparison score, the probability density of the modeled same-source distribution is divided by that of the different-source distribution. The histograms and modeled distributions, and the corresponding calculated LRs for both the 2D and 3D data are shown in Figure 2. The 95% bootstrap confidence intervals for the calculated LRs, based on 2000 bootstrap samples, are also shown to illustrate the sensitivity of the calculated LRs to the sampling (sampling error was not taken into account in the calculation of the LRs). While sampling for the different-source distributions we only used the comparison scores resulting from the two first test shots per firearm. This was done to correct for dependency between comparison scores. Our comparison score was a similarity-only score that did not take into account the typicality of the features (see Morrison and Enzinger (2018) [66] for more information about possible effects hereof on the calculated LRs).

To enable a robust calculation of the LRs it is necessary that there is sufficient empirical data. Because of insufficient data in the tails of the modeled distributions the calculated LRs become sensitive to sampling error, which is visualized by the large bootstrap confidence intervals for the LRs based on the tail areas of the histograms (Figure 2). A straightforward approach to decide which LRs can be considered to be robust with regard to sampling error is to limit the minimum and maximum LR values based on the size of the used sample set used ([67]). As a rule of thumb, the LR should not be smaller than 1 divided by the number of same-source comparisons or larger than the number of different-source comparisons. The calculated LRs outside that range are increasingly based on extrapolation of the distributions with limited to no empirical underlying data and could be replaced by the conservative LR values at the bounds. Because of the number of our same-source (N = 200) and different-source (N = 79600) comparisons, the robust LRs will be between approximately

10^{-2} and 10^4 to 10^5 (see Vergeer et al. (2016) [67] and Morrison and Poh (2018) [68] for detailed suggestions to avoid overstating the LR).

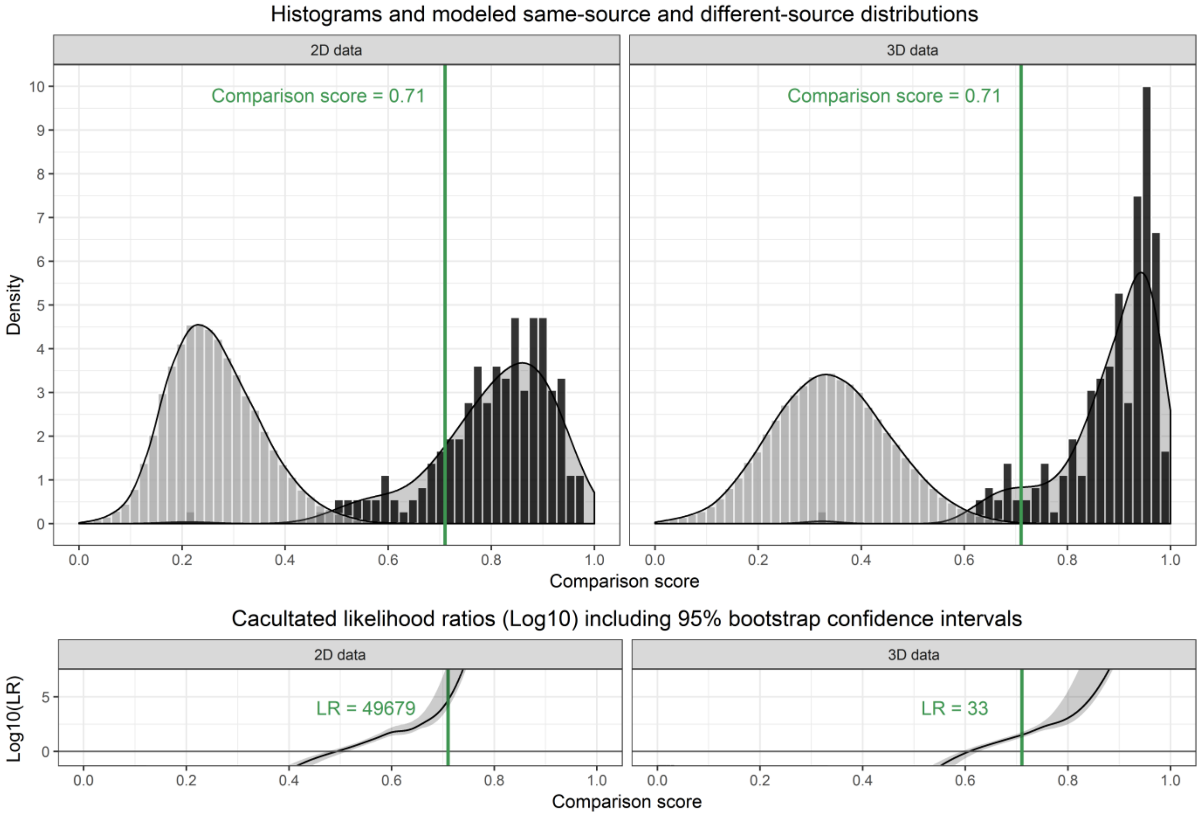

Figure 2

The histograms and modeled same-source (black bars) and difference-source distributions (gray bars) (top) and the corresponding calculated LRs and 95% bootstrap confidence intervals (below) for the 2D (left) and 3D (right) data. The vertical lines represent an arbitrarily chosen comparison score of 0.71 and the corresponding calculated LRs is indicated for the 2D and 3D data, respectively.

2.3 Examiner judgments

2.3.1 Participants

We invited forensic firearm examiners from Europe, North America, South America, Asia and Oceania by e-mail to participate and to extend the invitation to their direct colleagues. This e-mail contained a link to an online questionnaire that was used to acquire the examiner judgments. We asked each participant to provide some background information about their qualifications as a firearm examiner, whether they work for an accredited institute, their years of experience, their country of employment, and in what format they report their opinions in casework. A total of 112 recipients opened the online questionnaire and consented to the use of their data for this study. We use the data resulting from 77 of these participants for analysis. We excluded the data from the other subjects from further analysis because they were incomplete.

The 77 included participants were examiners from all invited continents. Of the participants, 75 stated that they were qualified examiners, one was in training and one taught firearm examination at the university, 56 worked for an accredited institute, and their years of experience ranged from 1 to 47 years (*M* = 16.3, *SD* = 8.9). Of the participants, 58 stated that they provide categorical conclusions in casework (i.e., exclusion / inconclusive / inclusion judgments), 13 provide probabilistic conclusions of whom 10 report likelihood ratios, and 6 apply the 5 step reporting scale as proposed in Pauw-Vugts et al. (2013) [31].

2.3.2 Study design

The online questionnaire started with a description of the purpose of the study. On the questionnaire the participants were shown a set of 60 comparison images. In these comparison images the 2D measurements of the striation patterns of two cartridge cases were aligned in correspondence to the alignment by the computer-based method and visualized in a way that is familiar to firearm examiners (side-by-side). Before being shown the 60 comparison images to be judged, the participants were shown an overview of comparison images showing features with various degrees of similarity (see Figure 3).

After that, the participants were asked to judge the degree of similarity of the aligned striation patterns on a scale consisting of : 1) (almost) no similarity, 2) a low degree of similarity, 3) a medium degree of similarity, 4) a high degree of similarity, and 5) almost total similarity.

After judging the degrees of similarity, the same 60 comparison images were again shown to the participants. Three additional questions were asked per comparison image:

1) Does the comparison provide support for the striations in the cartridge cases being the result of firing the cartridge cases with one or with two Glock pistols?

2) What is your judged degree of support for this conclusion?

3) Would you have provided an inconclusive conclusion in casework?

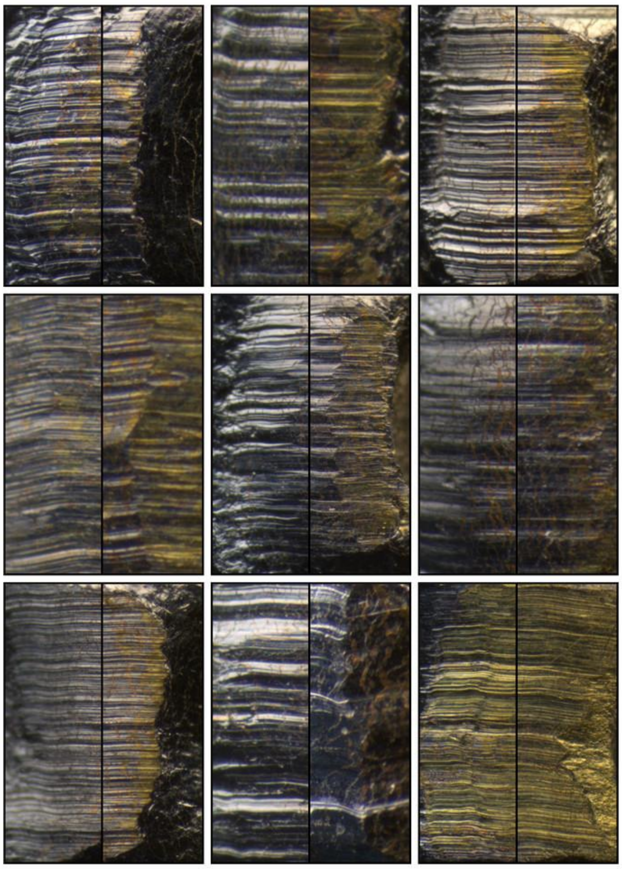

Figure 3.

Overview of nine comparison images showing features with various degrees of similarity. The comparisons were selected from the range of comparison scores resulting from the computer-based method. These comparison images were not part of the 60 comparison images used to acquire the examiner judgments.

The first question resulted in a source choice were the participants judged whether the comparison image provides support for two cartridge cases having been fired with one Glock pistol (same source) or with two Glock pistols (different source). For the second question the participants were asked to judge the degree of support for their source judgment on a six-step verbal scale (Table 1). The participants who stated that they report their results as an LR in casework received the same verbal scale, but defined by numerical frequency of occurrence ranges (e.g., moderate support (1 in 2 to 1 in 100 test fires). For same-source judgments participants were asked to assume the opposite, that the cartridge case on the right was actually fired with another Glock pistol. Keeping that assumption in mind they were asked in every how many test shots from other Glock pistols (Column 2) they would expect to find the same striation pattern (and resulting degree of similarity). For different-source judgments participants were again asked to assume the opposite, that the cartridge case on the right was actually fired with the same Glock pistol as the one on the left. Keeping that assumption in mind they were asked for how many test shots from that Glock pistol (Column 2) they would expect to find the same striation pattern (and resulting low degree of similarity) as seen in the cartridge case on the right. These judged frequencies of occurrence were used to approximate the resulting LRs as a random match equivalent [69] (Table 1). For same-source judgments, these pseudo LRs are calculated by setting the likelihood of the degree of similarity given the same-source proposition to 1 and dividing that by the judged relative frequency of occurrence of finding the same striation pattern (and resulting degree of similarity) with test shots from other Glock pistols (different-source proposition). For different-source judgments, the pseudo LRs are calculated by dividing the judged relative frequency of occurrence of finding the same striation pattern (and resulting degree of similarity) when the cartridge cases were fired with the same firearm (same-source proposition) by the likelihood of the degree of similarity given the different-source proposition, which was set to 1. The choice was made to use these approximated LRs because asking the examiners to provide the likelihood of the degrees of similarity given each of the propositions would have resulted in a (too) complex questionnaire, increasing the likelihood of misunderstanding of the task and distortion of the results.

The third question was whether participants would feel confident to report their judgment about the source in casework or would provide an ‘inconclusive’ conclusion. The order in which the comparison images were shown was randomized for each participant and when judging the degree of similarity and judging the source and degree of support per participant.

The 60 comparisons were selected from the 79,800 same-source and different-source comparisons as performed by the computer-based method and included 38 same-source and 22 different-source comparisons. The proportion of same-source comparisons was chosen to be higher as these were thought to occur more often in one-to-one comparisons in casework. The comparisons were selected mainly based on the calculated LRs for the 2D measurements reached with the computer-based method, while also considering the calculated LRs for the 3D measurements. The same-source and between-source distributions, necessary to calculate the LR of a comparison, were modeled for each comparison of two cartridge cases without including the test shots of the corresponding firearm(s).

Table 1

The used scale to judge the degree of support for the same-source or different-source judgments (Column 1). The subjects that report LRs in casework were also shown the information in Column 2. The approximated LRs based on random match equivalents for the same-source and different-source judgments are shown in Column 3 and 4, respectively.

Degree of support	Judged occurrence	Approximated LR same-source judgment	Approximated LR different-source judgments
Weak support	In 1 in 2 TO 1 in 10 test shots	2-10	0.1-0.5
Moderate support	In 1 in 10 TO 1 in 100 test shots	10-100	0.01-0.1
Moderately strong support	In 1 in 100 TO 1 in 1,000 test shots	100-1,000	0.001-0.01
Strong support	In 1 in 1,000 TO 1 in 10,000 test shots	1,000-10,000	0.0001-0.001
Very strong support	In 1 in 10,000 TO 1 in 1,000,000 test shots	10,000-1,000,000	0.000001-0.0001
Extremely strong support	In less than 1 in 1,000,000 test shots	>1,000,000	<0.000001

We put more emphasis on the 2D measurements because that would ensure that the same visual data would be available for the participants and the computer-based method. Comparisons were selected along the complete range of calculated LRs, including the calculated LRs which could be considered to be less robust (see the LR calculation subsection of the Materials and Methods section). We selected a relatively large proportion of low LRs for same-source comparison and high LRs for different-source comparisons to ensure an equal distribution of comparisons with varying LRs in the test set. Furthermore, we included all three same-source comparisons where the calculated LR (based on 2D and/or 3D measurement) was smaller than 1 and ten different-source comparisons where the calculated LR was larger than 1 (misleading evidence). This resulted in an overrepresentation of ‘difficult’ comparisons when compared to the available population of comparisons. Because the computer-based method is still experimental, we chose to also consider the calculated LRs based on the 3D data when selecting comparisons. When possible, we selected those comparisons that had fairly consistent calculated LRs based on 2D and 3D measurements. This was done to more robustly select comparisons which can be considered to be easier or harder. We considered the difference in degree of support steps (Table 1) between the 2D and 3D calculated LRs as a measure of consistency between LRs. By selecting comparison sets in this way we ensured that our test set contained samples that replicate the full range of comparison difficulty (as suggested by e.g., AAAS (2017) [70]). A list of the 60 selected comparisons with calculated LRs is shown in Appendix 1.

2.4 Analyses

Here we introduce the analyses that we perform to assess the validity and reliability of source judgments by examiners and the validity of the computer-based method, to determine the relation between examiners' judgments and the outcomes of the computer-based method, and to determine how calibrated the judged degrees of support of firearm examiners are.

2.4.1 Validity of source choices

To determine the validity of the outcomes of both the computer-based method and all the examiners combined we calculate the true and false positive rate, and the true and false negative rate based on the chosen same-source or different-source proposition. The true positive rate is also known as sensitivity and the true negative rate as specificity. The false positive and false negative rates relate to the rates of misleading evidence for different-source and same-source comparisons, respectively.

2.4.1.1 Computer-based method

For the validity analysis of the computer-based method a calculated LR above 1 is considered to be a true positive when the compared cartridge cases were indeed fired with one firearm (same source) and a false positive when they were fired with two firearms (different source). Likewise, a calculated LR below 1 is considered to be a true negative when the compared cartridge cases were indeed fired with two firearms and a false negative when they were fired with one firearm.

2.4.1.2 Firearm examiners

For the examiners, we compare the proposition for which they found support (from here on: source choice) to the known ground truth of the compared cartridge cases. A same-source choice is considered to be a true positive when the cartridge cases were indeed fired with one firearm and a false positive when they were fired with two firearms. Likewise, a different-source choice is considered to be a true negative when the cartridge cases were indeed fired with two firearms and a false negative when they were fired with one firearm.

We determine the validity of the source choices based on all 60 judged comparisons, whether or not the examiners did not feel confident to report these in casework ('inconclusives') and on only the source choices that the examiners felt confident to report, thus excluding the choices that the examiners judged to be 'inconclusive'. For all 60 judged comparisons we will also break down this analysis based on the judged degrees of support.

2.4.2 Reliability of examiner judgments

We study both the within-subject and between-subject reliability.

The within-subject reliability can be studied because the examiners judged both the degree of similarity and the degree of support for the same 60 comparisons. The degree of similarity is used by the examiners to provide a judgment about the source of the compared cartridge cases and about the degree of support for that judged source. To determine the within-subject reliability, we calculate the Spearman correlations between the judged degree of similarity and the judged degree of support per examiner, both for the same-source and different-source comparisons.

For the between-subject reliability we consider both the judged degree of similarity and the degree of support. We calculate the Spearman correlations between examiners for these two types of judgments, both for the same-source and different-source comparisons.

2.4.3 Relation between examiners' judgments and the outcomes of the computer-based method

To determine the relation between the examiners' judgments and the outcomes of the computer-based method we calculate the Spearman correlation for each examiner between the judged degree of similarity and the comparison scores based on the 2D measurements and between the judged degree of support and these comparison scores. We only consider the comparison scores based on the 2D measurements as this ensures that the same visual data was available for the examiners and the computer-based method.

2.4.4 Calibration of judged degrees of support

We look into the calibration of judged degrees of support because it has been argued that examiners are able to provide meaningful judgments of the degree of support for same-source versus different-source propositions based on their experience [56, 71].

Studying whether the judged degrees of support are calibrated is only possible for the examiners who report likelihood ratios. Their judged degrees of support were defined by numerical ranges, ensuring a similar interpretation of the verbal scale. The other examiners provided their judgments about the degree of support on a verbal scale. This verbal scale was not defined by numerical ranges and as a result the perception of the degrees of support will vary between examiners [72, 73] and this variability will probably be larger than for expressions defined by numerical ranges [72]. As a result, it is not meaningful to combine the judgments of those examiners.

We first aggregate the degree of support judgments of the examiners who report likelihood ratios based on the chosen proposition (same-source or different-source) and degree of support (weak support to extremely strong support). For these judgments the proportion of misleading choices is calculated for each degree of support.

To see whether the judged degrees of support are calibrated we compared the calculated proportions of misleading choices with the expected ranges of the proportions of misleading choices based on the ranges of the judged degrees of support. When judgments of the degree of support are well-calibrated, they are expected to fall within these ranges. These ranges are based on the approximated LR ranges resulting from the judged degrees of support, using Equation 1 [74]:

$$\frac{1}{\left(\frac{N_{same\ source}}{N_{different\ source}} \cdot LR\right)+1} = \frac{1}{\left(\frac{38}{22} \cdot LR\right)+1} \qquad \text{(Eq. 1)}$$

Furthermore, we calculate the Pearson correlation between the proportion of misleading choices and the expected ranges of the proportions of misleading choices based on the ranges of the judged degrees of support.

3 Results

3.1 Validity of source choices

The number of true positive, true negative, false positive and false negative source choices and the resulting true positive and true negative rates and the false positive and negative rates for the computer-based method and the examiners as well as the number of ‘inconclusive’ judgments for the examiners are shown in the confusion matrices in Table 2.

Table 2

Confusion matrices for the outcomes of the computer-based methods and the judgments of the examiners in relation to the ground truth of the comparison (same-source (SS) or different-source (DS) comparison). To calculate the True Positive Rates (TPR), True Negative Rates (TNR), False Positive Rates (FPR) and False Negative Rates (FNR) the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) conclusions were entered in the following equations: TPR = TP/(TP+TN), TNR = TN/(TN+FP), FPR = FP/(FP+TN) and FNR = FN/(FN+TP). The number of ‘inconclusive’ judgments are shown in the last matrix.

Computer-based method – 2D data – All data

N = 79800	SS comparison	DS comparison	
SS outcome (LR > 1)	198	1012	FPR = .013
DS outcome (LR < 1)	2	78588	FNR = .010
	TPR = .990	TNR = .987	

Computer-based method – 3D data – All data

N = 79800	SS comparison	DS comparison	
SS outcome (LR > 1)	198	999	FPR = .013
DS outcome (LR < 1)	2	78601	FNR = .010
	TPR = .990	TNR = .987	

Computer-based method – 2D data – Sixty selected comparisons

N = 60	SS comparison	DS comparison	
SS outcome (LR > 1)	36	10	FPR = .455
DS outcome (LR < 1)	2	12	FNR = .053
	TPR = .947	TNR = .545	

Computer based-method – 3D data – Sixty selected comparisons

N = 60	SS comparison	DS comparison	
SS outcome (LR > 1)	36	5	FPR = .227
DS outcome (LR < 1)	2	17	FNR = .053
	TPR = .947	TNR = .773	

Examiners – All sixty selected comparisons

N = 4620	SS comparison	DS comparison	
SS judgment	2726	322	FPR = .190
DS judgment	200	1372	FNR = .068
	TPR = .932	TNR = .810	

Examiners – Selected comparisons excluding the ‘inconclusive’ judgments

N = 3318	SS comparison	DS comparison	
SS judgment	2365	95	FPR = .108
DS judgment	74	784	FNR = .030
	TPR = .970	TNR = .892	

Examiners – Number of 'inconclusive' judgments

N = 1302	SS comparison	DS comparison
SS judgment	361	227
DS judgment	126	588

When taking into account the calculated LRs of all 79,800 comparisons, the true positive rates (sensitivity) and the true negative rates (specificity) of the computer-based method were high, at 99.0% and 98.7% for both the 2D and 3D measurements, while the complementary false negative and the false positive rates were low. The true positive rates of the 60 selected comparisons were slightly lower, at 94.7% for the 2D and 3D measurements, but the true negative rates were a lot lower, at 54.5% and 77.3% for the 2D and 3D measurements, respectively. These lower true negative rates are to be expected with the selection of a ‘difficult’ comparison set. This difference is mainly caused by the selection of a relatively large proportion of comparisons with misleading same-source LRs.

The true positive rate of the examiners for the 60 comparisons is slightly lower (93.2%) than that of the computer-based method, while the true negative rate is higher (81.0%). Based on these results, the examiners seem to be slightly less proficient at identifying same-source comparisons correctly, while they are better at identifying different-source comparisons correctly.

When we exclude the judgments that the examiner did not feel confident to report in casework from the analysis (considering the 'inconclusive' judgments), we see that the true positive rate and the true negative rate are slightly higher (97.0% and 98.2%, respectively) when compared to all judgments. When only considering the judgments that the examiners felt confident to report, the validity of the examiners' judgments is (slightly) higher than that of the computer-based method on both same-source and different-source comparisons.

Even though the examiners as a group seem to provide quite valid judgments there are large individual differences (see Figure 4 and Table 3).

When we break down the analysis based on judged degree of support, we see that the validity of the examiner judgments increased with increasing judged degree of support for a source proposition (Table 4). This is shown by the increasing true positive and true negative rates with increasing judged degree of support for a source proposition.

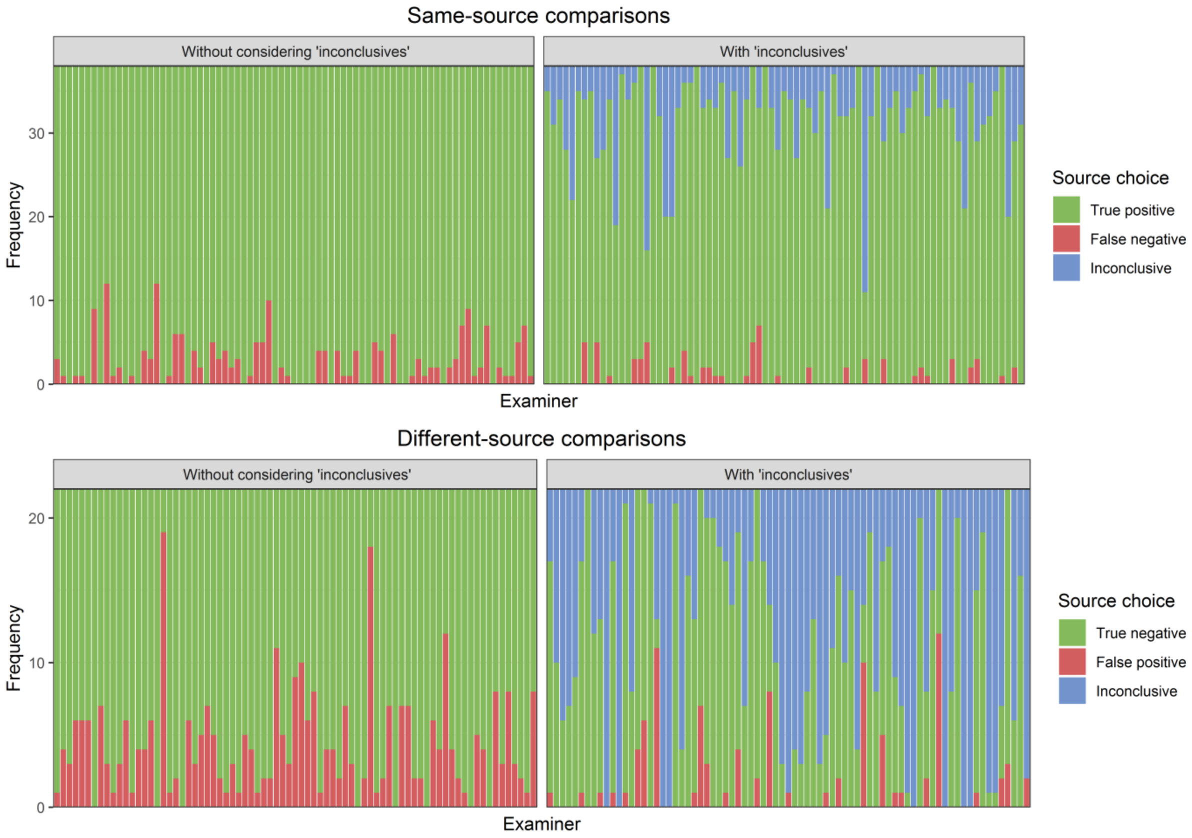

Figure 4

The number of true positive, false positive, true negative, and false negative choices, and the number of inconclusives per examiner, for the same-source comparison (top), different-source comparisons (bottom), without considering the inconclusives (left) and with inconclusives (right).

Table 3

Specification of the means, standard deviations, and 95% confidence intervals for the True Positive and the True Negative Rates (TPR and TNR), and the False Positive and the False Negative Rates (FPR and FNR) for all of the 77 examiners' source choices and for those which they felt confident to report (i.e. excluding the inconclusives).

Validity metric		TPR	FPR	TNR	FNR
Formula		*TP/(TP+FN)*	*FP/(FP+TN)*	*TN/TN+FP)*	*FN/(FN+TP)*
All source choices	*M*	.932	.190	.810	.068
	SD	.077	.167	.167	.077
	95%-CI	[.915, .949]	[.153, .227]	[.773, .847]	[.051, .085]
Excluding inconclusives	*M*	.966	.115	.885	.034
	SD	.063	.230	.230	.063
	95%-CI	[.952, .980]	[.064, .167]	[.833, .936]	[.020, .048]

Table 4

Confusion matrices for the judgments of the examiners in relation to the ground truth of the comparison (same-source (SS) or different-source (DS) comparison) per judged degree of support.

Weak support

N = 504	SS comparison	DS comparison	
SS outcome (LR > 1)	84	153	FPR = .415
DS outcome (LR < 1)	51	216	FNR = .378
	TPR = .622	TNR = .585	

Moderate support

N = 696	SS comparison	DS comparison	
SS outcome (LR > 1)	240	105	FPR = .264
DS outcome (LR < 1)	58	293	FNR = .195
	TPR = .805	TNR = .736	

Moderately strong support

N = 689	SS comparison	DS comparison	
SS outcome (LR > 1)	351	46	FPR = .151
DS outcome (LR < 1)	34	258	FNR = .088
	TPR = .912	TNR = .849	

Strong support

N = 1009	SS comparison	DS comparison	
SS outcome (LR > 1)	664	15	FPR = .048
DS outcome (LR < 1)	35	295	FNR = .050
	TPR = .950	TNR = .952	

Very strong support

N = 1070	SS comparison	DS comparison	
SS judgment	826	3	FPR = .013
DS judgment	17	224	FNR = .020
	TPR = .980	TNR = .987	

Extremely strong support

N = 652	SS comparison	DS comparison	
SS judgment	561	0	FPR = 0
DS judgment	5	86	FNR = .009
	TPR = .991	TNR = 1	

3.2 Within-subject reliability of judgments

The summary of the Spearman correlations between the judged degree of similarity and the judged degree of support per examiner is shown in Table 5, both for the same-source and different-source comparisons.

Table 5

The mean, standard deviation, and 95% confidence interval of the Spearman correlations per examiner between the judged degree of similarity and degree of support, for the same-source and different-source comparisons.

Source of comparisons	*N*	*M*	*SD*	*95%- CI*
Same source	77	.708	.100	[.686, .731]
Different source	77	.518	.221	[.467, .568]

3.3 Between-subject reliability of judgments

The summary of the Spearman correlations between examiners for the judged degree of similarity and the judged degree of support is shown in Table 6, both for the same-source and different-source comparisons. Comparing the judgments of all examiners results in a total of 2926 (77×76/2) correlations per combination of comparison source (same-source or different-source comparison) and type of judgment (degree of similarity or degree of support). The judgments of some of the examiners for the different-source comparisons showed no variability, decreasing the total number of between-subject correlations.

Table 6

The sample sizes, means, standard deviations, and 95% confidence intervals of the Spearman correlations between examiners for the judged degree of similarity and degree of support, for the same-source and different-source comparisons.

Source of comparisons	Degree of similarity				Degree of support			
	N	*M*	*SD*	*95%-CI*	*N*	*M*	*SD*	*95%-CI*
Same source	2926	.620	.111	[.616, .624]	2926	.631	.137	[.626, .636]
Different source	2775	.395	.208	[.387, .402]	2850	.457	.210	[.449, .465]

3.4 Relation between examiners' judgments and the outcomes of the computer-based method

The summary of the Spearman correlations between the examiners' judgments (degree of similarity and degree of support) and the outcomes of the computer-based method based on the 2D measurements is shown in Table 7, both for the same-source and different-source comparisons.

Table 7

The mean, standard deviation, and 95% confidence interval of the Spearman correlations between the judged degree of similarity and the computer-based method's comparison scores based on the 2D measurements and between the judged degree of support and the comparison scores, for the same-source and different-source comparisons.

Type of judgment and source of comparisons		*N*	*M*	*SD*	*95%-CI*
Degree of similarity	Same source	77	.368	.124	[.340, .395]
	Different source	75	.492	.190	[.450, .535]
Degree of support	Same source	77	.375	.135	[.345, .405]
	Different source	76	.509	.184	[.468, .551]

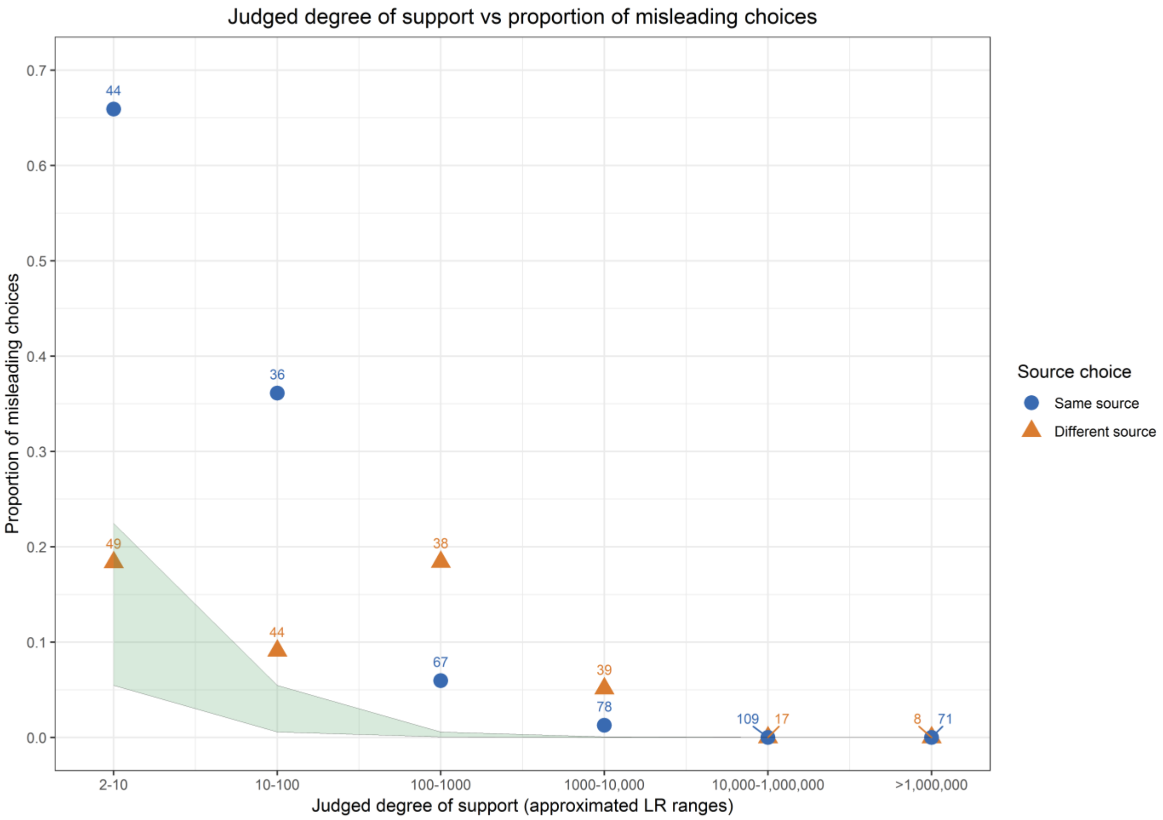

Figure 5

The proportion of misleading choices (misleading same-source choices (circles) and misleading different-source choices (triangles) per combination of chosen source and judged degree of support, including the number of judgments per group. The shaded area represents the area in which the proportion of misleading choices should lie if the judged degrees of support are well-calibrated. For the different-source choices the approximated LR ranges for the reversed propositions order are shown (1/LR).

3.5 Calibration of judged degrees of support

There were 10 examiners who reported likelihood ratios. The proportion of misleading choices per judged degree of support is shown in Figure 5. In the same figure the expected ranges of misleading choices given the judged degrees of support are shown (shaded area). These were calculated using Equation 1.

The Pearson correlation indicates a significant positive association between the actual proportion of misleading same-source choices and the upper bound (*r*(4) = .957, *p* = .003) and lower bound (*r*(4) = .907, *p* = .012) of the expected proportion of misleading choices for the same-source choices. For the different-source choices the Pearson correlations are lower and not significant (*r*(4) = .615, *p* = .194 for the upper bounds and *r*(4) = .596, *p* = .211 for the lower bounds). Although the positive associations of the same-source choices are large, the judged degrees of support are not well-calibrated as overall the actual proportion of misleading choices is (much) higher than would be expected based on the judged degrees of support (Figure 5).

4 Discussion

Following a shooting incident, firearm examiners are asked to judge whether e.g., two cartridge cases were fired with the same or different firearms, or whether they were fired with a submitted firearm or not. The validity of such judgments is increasingly questioned [1, 8]. We believe that the results of this study show that, although there are individual differences (Figure 4 and Table 3), the true positive rates (sensitivity) and the true negative rates (specificity) of the examiners were quite high. At the same time, the complementary false negative and the false positive rates were rather low (Table 2). When looking into the individual results we see that percentages of false positive choices (Figure 4, bottom left) are fairly high. This might be related to our choice to select ‘difficult’ comparisons, which we will discuss later. When comparing the results of the examiners to the used computer-based method, the examiners seem to be slightly less proficient at identifying same-source comparisons correctly, while they are better at identifying different-source comparisons correctly. The result that the examiners are better at identifying different-source comparisons correctly, corresponds with that of another study, focusing on the comparison of striation patterns from screwdrivers, where the examiners also outperform the computer-based method [75].

The source choices of the examiners were more valid when the choices that the examiners did not feel confident to report in casework ('inconclusives') were excluded from the analysis (Table 2). Compared to the results of the computer-based method the examiners' validity is then (slightly) higher for both same-source and different-source comparisons. Allowing examiners to judge a comparison as 'inconclusive' fitted with the current practice of most participants, who provide categorical conclusions in casework (e.g., exclusion / inconclusive / inclusion judgments). At the same time, this allowed the examiners to control the difficulty of the comparison set on which the validity analysis would be performed. The examiners had the liberty to judge each 'difficult' comparison as 'inconclusive', while the computer-based method did not get that liberty. Because of this, we argue that the first comparison between examiners and the computer-based method, including examiner judgment on all 60 comparisons, is most fair.

The result that the source choices of the examiners are quite valid and specifically that the true negative rate is higher than that of the computer-based method on the set of 60 comparisons, substantiates their expertise. This does not mean that the computer-based method is not useful and cannot be improved upon. The computer-based method is also quite valid and can easily deal with a large number of comparisons, while that will be far more time consuming for the examiners. Because of the latter capability we chose to select the set of 60 comparisons based on the outcomes of the computer-based method. This resulted in a comparison set which over-represents 'difficult' comparisons for the computer-based method. Because both the examiners and the computer-based method consider similar features and consider a degree of similarity we argue that this was also an approximately equally difficult set for the examiners. A way to test this assumption, which we did not deem feasible, could be to reverse the order of events, i.e. by first letting the examiners perform all possible 79,800 comparisons, and to then select a 'difficult test set' to provide to the computer-based method. Additional studies to improve the validity of the, at this moment, experimental computer-based method could potentially decrease the overlap between same-source and different-source comparison scores (Figure 2) and consequently increase its performance.

For several reasons it is not possible to directly relate the true positive and true negative rates, and the false positive and negative rates of the examiners from this study to casework. One of these reasons is that the 60 comparisons we used were selected to over-represent 'difficult' comparisons. In addition, the use of the online questionnaire did not enable the examiners to manually compare the features of the cartridge cases as they would normally do in casework. They could not include in their considerations the features of other firearm components, and their results and conclusions were not peer reviewed. Enabling examiners to follow their standard operating procedures could result in better performance.

The judged degree of support for a source proposition can be considered as a measure of difficulty of the comparison. A higher judged degree of support would correspond to a lower comparison difficulty and vice versa. When considering our results as such, it can be seen that the true positive and true negative rates increased (increasing validity of the source choices of the examiners) with increasing judged degree of support (decreasing difficulty). This result shows that trying to provide one overall error-rate for a forensic discipline, as is implied in the PCAST report [8], is counter-productive. The expected rate of misleading evidence will depend on the combination of the examiner's expertise and the difficulty of the specific comparison. Reporting judged evidential strength in probabilistic terms, such as a likelihood ratio, instead of a categorical conclusion (i.e., exclusion / inconclusive / inclusion judgments) would enable examiners to provide the courts with a more informative conclusion.

The strong positive association between judgments of the degree of similarity and degree of support per examiner (Table 5) shows that the within-subject reliability is quite high for same-source comparisons and slightly lower for different-source comparisons. For the between-subject reliability of both the judgments about degree of similarity and degree of support a similar difference is seen between same-source and different-source comparisons (Table 6). The between-subject reliability is lower than the within-subject reliability, but still shows a strong positive association between the examiners for same-source judgments. The moderate to high within- and between-subject reliability provides support for the proposition that the examiners generally reach similar conclusions.

The moderate positive associations between the judged degrees of similarity and judged degrees of support by the examiners, and the comparison scores by the computer-based method (for the 2D measurements) do not provide strong support for their coherence. Although both the examiners and the computer-based method take into account the degree of similarity of striation patterns, they might do so differently. Future work could focus on the exploration of the applied (cognitive) mechanisms of the examiners and the computer-based method to establish the strengths and weaknesses of the two. Such information could assist in the exploration of the possibilities of combining the judgments of examiners and the outcomes of the computer-based method to increase the overall validity.

At this moment, a large-scale implementation of computer-based methods for the comparison of striation and impression patterns resulting from firearms seems to be years away. This mainly has to do with the need to set up sizable reference databases for the evaluation of comparison scores. For a thorough evaluation of the evidence, such reference databases will eventually be necessary for the numerous types of marks that can be compared. Among others, these could include databases on impression patterns resulting from a firearm's breechface, firing pin or ejector and on striation patterns resulting from a firearm's barrel or firing pin aperture. It is inconceivable that such databases will need to be set up for various manufacturing processes or manufacturers. Furthermore, the acquisition parameters and the comparison algorithms and their parameters will need to be optimized and the methods should be thoroughly validated (for a proposed guideline for validation see Meuwly, Ramos and Haraksim (2017) [76]). This does not mean that a smaller scale implementation of computer-based methods in casework cannot be achieved in a shorter timespan for specific firearm marks and manufacturing processes or manufacturers. This study's aim of comparing the results of examiners to a computer-based method is a first step towards combining the outcomes of both in casework. Because we did not find a strong correlation between the examiner judgments and the outcomes of the computer-based method it seems plausible that their conclusions will also differ in casework. Montani, Marquis, Egli Anthonioz and Champod (2019) [77] explore a way to reconcile differing conclusions. They suggest that in a forensic report, the outcomes of a computer-based method should carry more weight than the judgments of examiners when both consider the same features and when the computer-based method is validated and applied within the defined boundaries of usage. The reasoning behind this suggestion is that the computer-based method offers systematic measures and as a result has higher scientific credentials than examiner judgments. Examiners would subsequently be allowed to adapt a computer-based method's conclusion when they consider additional characteristics that are not incorporated in the computer-based method.

When considering the degree of support judgments of the 10 examiners who report LRs in casework it is possible to study how calibrated their judgments are. The actual proportions of misleading source choices are much higher than the expected ranges of misleading choices given the judged degrees of support, showing that the judged degrees of support are not well-calibrated (Figure 5). This effect is most clear for the same-source choices, where there is a significant and strong positive association between the actual proportion of misleading choices and the expected ranges of misleading choices. These expected ranges are based on the approximated LRs. This means that the examiners have judged frequencies of occurrence that are lower than warranted by the proportions of misleading choices, resulting in an overestimation of the degree of support. Because worldwide most examiners do not report LRs and because the majority of the examiners who report LRs in this study are employed in one country care should be taken when generalizing this result to a broader population of forensic (firearm) examiners.

The result that the judged degrees of support of the 10 examiners who report LRs in casework are not well-calibrated does not fully corroborate the assumption that examiners are able to provide meaningful judgments of the probabilities of features when they provide judgments about the degree of support for a source judgments based on their experience [56, 71]. The result does correspond with the results of other studies on the calibration of judgments [78-80]. These studies show that many expert populations show over-extremity in their judgments [78], meaning that their probability estimates lie too close to 0 or 1. When we look at the actual proportion of misleading choices, the examiners judged lower relative frequencies of occurrence (and thus more extreme LRs) than expected if their judgments would have been well-calibrated. This can be seen as overconfidence, where examiners provide unwarranted support for either same-source or different-source propositions, resulting in LRs that are too high or too low, respectively. Because such examiner judgments are relied upon by the judicial systems it is important that they are well-calibrated. Simply warning examiners about overconfidence [81] or asking them to explain their judgments [82] does not necessarily decrease overconfidence of judgments. Providing performance feedback to examiners, a necessary component of ‘deliberate practice’ to acquire expertise [83], does seem to reduce overconfidence and increase calibration [78, 84-87].

We chose to only define the used verbal degree of support scale by numerical frequency of occurrence ranges for the examiners that report likelihood ratios in casework. We made this choice because we reasoned that asking examiners who had no experience with this to assess these frequency of occurrence ranges could result in unreliable judgments. Future studies could test whether examiners that report likelihood ratios are more proficient in judging these frequency of occurrence ranges than examiners who provide e.g., categorical conclusion. Furthermore, research on the use of verbal and numerical expressions to elicit and receive probability judgments has shown that peoples’ preferences vary. One third prefers to receive and express information numerically, one third prefers verbal expressions for both, and one third prefers to receive information numerically and to express it verbally to convey the imprecision of their judgments [88]. This latter preference has been dubbed the 'communication mode preference paradox' [89]. Verbal expressions of probabilities are perceived to convey the imprecision of judgments better than numerical expressions and they are perceived to be more natural, and easier to communicate and express [90]. At the same time, the within and between variability is smaller when numerical probability expressions are used instead of verbal expressions [72]. To facilitate a consistent interpretation of probability expressions it is advised to provide a short scale of standardized verbal expressions [73, 91, 92], with a pre-defined rank-order [93, 94] and to define these by (fixed ranges of) numerical probabilities [93-98]. Our used degree of support scale, defined by numerical frequency of occurrence ranges fulfills these criteria. In current casework practice these criteria seem to be met by examiners that report their judgment as a likelihood ratio, defining verbal expression of the degree of support by numerical ranges [e.g., 53, 54, 56].

5 Conclusions

We conclude that the true positive rates (sensitivity) and the true negative rates (specificity) of firearm examiners are quite high and that they generally reach similar conclusions. The examiners seem to be slightly less proficient at identifying same-source comparisons correctly, while they outperform the used computer-based method at identifying different-source comparisons. At the same time, the judged degrees of support for these source choices are not well-calibrated. This could be improved by implementing performance feedback to reduce overconfidence. Further work on the strengths and weaknesses of the examiners and the computer-based method could assist in the exploration of the possibilities of combining the judgments of examiners and the outcomes of computer-based method to increase the overall validity.

References

[1] Committee on Identifying the Needs of the Forensic Sciences Community: National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, DC, USA, 2009.

[2] I.E. Dror, S.A. Cole, The vision in "blind" justice: expert perception, judgment, and visual cognition in forensic pattern recognition, Psychonomic bulletin & review 17 (2010) 161-7.

[3] S.M. Kassin, I.E. Dror, J. Kukucka, The forensic confirmation bias: Problems, perspectives, and proposed solutions, Journal of Applied Research in Memory and Cognition 2 (2013) 42-52.

[4] 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, Science & Justice 56 (2016) 155-156.

[5] I.E. Dror, S.M. Kassin, J. Kukucka, New application of psychology to law: Improving forensic evidence and expert witness contributions, Journal of Applied Research in Memory and Cognition 2 (2013) 78-81.

[6] D.M. Risinger, M.J. Saks, W.C. Thompson, R. Rosenthal, The Daubert/Kumho Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion, California Law Review 90 (2002) 1-56.

[7] 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, Science & Justice 43 (2003) 77-90.

[8] Executive Office of the President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, 2016.

[9] J.L. Mnookin, S.A. Cole, I.E. Dror, B.A.J. Fisher, M.M. Houck, K. Inman, D.H. Kaye, J.J. Koehler, G. Langenburg, D.M. Risinger, N. Rudin, J. Siegel, D.A. Stoney, The need for a research culture in the forensic sciences, UCLA Law Review 58 (2011) 725-779.

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

[11] S.A. Cole, Implementing counter-measures against confirmation bias in forensic science, Journal of Applied Research in Memory and Cognition 2 (2013) 61-62.

[12] W.C. Thompson, Painting the target around the matching profile: the Texassharpshooter fallacy in forensic DNA interpretation, Law, Probability and Risk 8 (2009) 257-276.

[13] W.C. Thompson, What role should investigative facts play in the evaluation of scientific evidence?, Australian Journal of Forensic Sciences 43 (2011) 123-134.

[14] I.E. Dror, Practical Solutions to Cognitive and Human Factor Challenges in Forensic Science, Forensic Science Policy & Management: An International Journal 4 (2013) 105-113.

[15] D.E. Krane, S. Ford, J.R. Gilder, K. Inman, A. Jamieson, R. Koppl, I.L. Kornfield, D.M. Risinger, N. Rudin, M.S. Taylor, W.C. Thompson, Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation, Journal of forensic sciences 53 (2008) 1006-1007.

[16] I.E. Dror, W.C. Thompson, C.A. Meissner, I.L. Kornfield, D.E. Krane, M.J. Saks, D.M. Risinger, Letter to the Editor - Context Management Toolbox: A Linear Sequential Unmasking (LSU) Approach for Minimizing Cognitive Bias in Forensic Decision Making, Journal of forensic sciences 60 (2015) 1111-1112.

[17] 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, Science & Justice 56 (2016) 113-122.

[18] E.J.A.T. Mattijssen, R.D. Stoel, W. Kerkhoff, Minimizing Contextual Bias in Forensic Firearms Examinations, Wiley Encyclopedia of Forensic Science, 2015, pp. 1-7.

[19] B. Found, J. Ganas, The management of domain irrelevant context information in forensic handwriting examination casework, Science & Justice 53 (2013) 154-8.

[20] K.N. Ballantyne, G. Edmond, B. Found, Peer review in forensic science, Forensic science international 277 (2017) 66-76.

[21] G. Edmond, A. Towler, B. Growns, G. Ribeiro, B. Found, D. White, K. Ballantyne, R.A. Searston, M.B. Thompson, J.M. Tangen, R.I. Kemp, K. Martire, Thinking forensics: Cognitive science for forensic practitioners, Science & Justice 57 (2017) 144-154.

[22] N.K.P. Osborne, M.C. Taylor, Contextual information management: An example of independent-checking in the review of laboratory-based bloodstain pattern analysis, Science & Justice 58 (2018) 226-231.

[23] E.J.A.T. Mattijssen, C.L.M. Witteman, C.E.H. Berger, R.D. Stoel, Cognitive Biases in the Peer Review of Bullet and Cartridge Case Comparison Casework: A Field Study, Science & Justice 60 (2020) 337-346.

[24] T.G. Fadul, G.A. Hernandez, S. Stoiloff, S. Gulati, An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides, Miami-Dade Police Department Crime Laboratory, 2011.

[25] T.G. Fadul, G.A. Hernandez, E. Wilson, S. Stoiloff, S. Gulati, An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing consecutively manufactured Glock EBIS barrels with the same EBIS pattern, 2013.

[26] J.E. Hamby, S. Norris, N.D. Petraco, Evaluation of GLOCK 9 mm Firing Pin Aperture Shear Mark Individuality Based On 1,632 Different Pistols by Traditional Pattern Matching and IBIS Pattern Recognition, Journal of forensic sciences 61 (2016) 170-176.

[27] E.D. Smith, Cartridge Case and Bullet Comparison Validation Study with Firearms Submitted in Casework, AFTE Journal 37 (2005) 130-135.

[28] J.E. Hamby, J. Brundage, J.W. Thorpe, The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries, AFTE Journal 42 (2009) 99-110.

[29] M.A. Keisler, Isolated Pairs Research Study, AFTE Journal 50 (2018) 56-58.

[30] M. Cazes, J. Goudeau, Validation Study Results from Hi-Point Consecutively Manufactured Slides, AFTE Journal 45 (2013) 175-177.

[31] P. Pauw-Vugts, A. Walters, L. Øren, L. Pfoser, FAID2009: Proficiency Test and Workshop, AFTE Journal 45 (2013) 115-127.

[32] W. Kerkhoff, R.D. Stoel, E.J.A.T. Mattijssen, C.E.H. Berger, F.W. Didden, J.H. Kerstholt, A part-declared blind testing program in firearms examination, Science & Justice 58 (2018) 258-263.

[33] F. Riva, C. Champod, Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases, Journal of forensic sciences 59 (2014) 637-647.

[34] X. Zheng, J. Soons, T.V. Vorburger, J. Song, T. Renegar, R. Thompson, Applications of surface metrology in firearm identification, Surface Topography: Metrology and Properties 2 (2014) 014012.

[35] S. Yammen, P. Muneesawang, Cartridge case image matching using effective correlation area based method, Forensic science international 229 (2013) 27-42.

[36] M. Tong, J. Song, W. Chu, R.M. Thompson, Fired Cartridge Case Identification Using Optical Images and the Congruent Matching Cells (CMC) Method, Journal of research of the National Institute of Standards and Technology 119 (2014) 575-582.

[37] U. Sakarya, O. Topçu, U.M. Leloğlu, M. Soysal, E. Tunali, Automated region segmentation on cartridge case base, Forensic science international 222 (2012) 277-287.

[38] J. Song, Proposed “Congruent Matching Cells (CMC)” Method for Ballistic Identification and Error Rate Estimation, AFTE Journal 47 (2015) 177-185.

[39] N.D.K. Petraco, L. Kuo, H. Chan, E. Phelps, C. Gambino, P. McLaughlin, F. Kammerman, P. Diaczuk, P. Shenkin, J.E. Hamby, Estimates of Striation Pattern Identification Error Rates by Algorithmic Methods, AFTE Journal 45 (2013) 235-244.

[40] C. Gambino, P. McLaughlin, L. Kuo, F. Kammerman, P. Shenkin, P. Diaczuk, N. Petraco, J. Hamby, N.D.K. Petraco, Forensic surface metrology: tool mark evidence, Scanning 33 (2011) 272-278.

[41] W. Chu, R.M. Thompson, J. Song, T.V. Vorburger, Automatic identification of bullet signatures based on consecutive matching striae (CMS) criteria, Forensic science international 231 (2013) 137-41.

[42] J. Song, T.V. Vorburger, W. Chu, J. Yen, J.A. Soons, D.B. Ott, N.F. Zhang, Estimating error rates for firearm evidence identifications in forensic science, Forensic science international 284 (2018) 15-32.

[43] Z. Chen, J. Song, W. Chu, J.A. Soons, X. Zhao, A convergence algorithm for correlation of breech face images based on the congruent matching cells (CMC) method, Forensic science international 280 (2017) 213-223.

[44] H. Zhang, J. Song, M. Tong, W. Chu, Correlation of firing pin impressions based on congruent matching cross-sections (CMX) method, Forensic science international 263 (2016) 186-193.

[45] X.H. Tai, W.F. Eddy, A Fully Automatic Method for Comparing Cartridge Case Images, Journal of forensic sciences 63 (2018) 440-448.

[46] H. Zhang, J. Gu, J. Chen, F. Sun, H. Wang, Pilot study of feature-based algorithm for breech face comparison, Forensic science international 286 (2018) 148-154.

[47] S. Bigdeli, H. Danandeh, M. Ebrahimi Moghaddam, A correlation based bullet identification method using empirical mode decomposition, Forensic science international 278 (2017) 351-360.

[48] F. Riva, R. Hermsen, E.J.A.T. Mattijssen, P. Pieper, C. Champod, Objective Evaluation of Subclass Characteristics on Breech Face Marks, Journal of forensic sciences 62 (2017) 417-422.

[49] E.F. Law, K.B. Morris, C.M. Jelsema, Determining the number of test fires needed to represent the variability present within 9mm Luger firearms, Forensic science international 276 (2017) 126-133.

[50] E.F. Law, K.B. Morris, C.M. Jelsema, Determining the number of test fires needed to represent the variability present within firearms of various calibers, Forensic science international 290 (2018) 56-61.

[51] C.C.G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd ed., John Wiley and Sons, Chichester, UK, 2004.

[52] Committee for the Advancement of the Science of Firearm & Toolmark Identification, Theory of Identifcation as it Relates to Toolmarks: Revised, AFTE Journal 43 (2011) 287.

[53] European Network of Forensic Science Institutes, ENFSI guideline for evaluative reporting in forensic science, 2016.

[54] S. Bunch, G. Wevers, Application of likelihood ratios for firearm and toolmark analysis, Science & Justice 53 (2013) 223-229.

[55] W. Kerkhoff, R.D. Stoel, E.J.A.T. Mattijssen, R. Hermsen, The Likelihood Ratio Approach in Cartridge Case and Bullet Comparison, AFTE Journal 45 (2013) 284-289.

[56] W. Kerkhoff, R.D. Stoel, E.J.A.T. Mattijssen, R. Hermsen, P. Hertzman, Đ. Hazard, M. Gallidabino, T. Hicks, C. Champod, Cartridge Case and Bullet Comparison: Examples of Evaluative Reporting, AFTE Journal 49 (2017) 111-121.

[57] Leica Microsystems GmbH, Leica FS C. <https://www.leica-microsystems.com/products/light-microscopes/p/leica-fs-c/>, 2019 (accessed 26 march, 2019.).

[58] Leica Microsystems GmbH, Leica DFC490. <https://www.leica-microsystems.com/products/microscope-cameras/details/product/leica-dfc490/>, 2019 (accessed 26 March, 2019.).

[59] Loci Forensics B.V., Forensic Sil. <https://www.lociforensics.nl/>, 2019 (accessed March 26, 2019.).

[60] Alicona GmbH, InfiniteFocusSL. <https://www.alicona.com/en/products/infinitefocussl/>, 2019 (accessed March 23, 2019.).

[61] M. Baiker, I. Keereweer, R. Pieterman, E. Vermeij, J. van der Weerd, P. Zoon, Quantitative comparison of striated toolmarks, Forensic science international 242 (2014) 186-199.

[62] M. Baiker, N.D.K. Petraco, C. Gambino, R. Pieterman, P. Shenkin, P. Zoon, Virtual and simulated striated toolmarks for forensic applications, Forensic science international 261 (2016) 43-52.

[63] M. Baiker, R. Pieterman, P. Zoon, Toolmark variability and quality depending on the fundamental parameters: Angle of attack, toolmark depth and substrate material, Forensic science international 251 (2015) 40-49.

[64] D.L. Garcia, R. Pieterman, M. Baiker, Influence of the axial rotation angle on tool mark striations, Forensic science international 279 (2017) 203-218.

[65] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018.

[66] G.S. Morrison, E. Enzinger, Score based procedures for the calculation of forensic likelihood ratios - Scores should take account of both similarity and typicality, Science & Justice 58 (2018) 47-58.

[67] P. Vergeer, A. van Es, A. de Jongh, I. Alberink, R. Stoel, Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating?, Science & Justice 56 (2016) 482-491.

[68] G.S. Morrison, N. Poh, Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors, Science & Justice 58 (2018) 200-218.

[69] W.C. Thompson, Discussion paper: Hard cases make bad law—reactions to R v T, Law, Probability and Risk 11 (2012) 347-359.

[70] AAAS, Forensic Science Assessments: A Quality and Gap Analysis - Latent Fingerprint Examination, in: W.C. Thompson, J.P. Black, A.K. Jain, J.B. Kadane (Eds.) 2017.

[71] A. Biedermann, P. Garbolino, F. Taroni, The subjectivist interpretation of probability and the problem of individualisation in forensic science, Science & Justice 53 (2013) 192-200.

[72] D.V. Budescu, S. Weinberg, T.S. Wallsten, Decisions based on numerically and verbally expressed uncertainties, Journal of Experimental Psychology: Human Perception and Performance 14 (1988) 281-294.

[73] D.V. Budescu, T.S. Wallsten, Consistency in interpretation of probabilistic phrases, Organizational Behavior and Human Decision Processes 36 (1985) 391-405.

[74] B. Robertson, G.A. Vignaux, C.E.H. Berger, Interpreting Evidence: Evaluating Forensic Science in the Courtroom, John Wiley & Sons, Chichester, UK, 2016, p. 92.

[75] L.S. Chumbley, M.D. Morris, M.J. Kreiser, C. Fisher, J. Craft, L.J. Genalo, S. Davis, D. Faden, J. Kidd, Validation of Tool Mark Comparisons Obtained Using a Quantitative, Comparative, Statistical Algorithm, Journal of forensic sciences 55 (2010) 953-961.

[76] D. Meuwly, D. Ramos, R. Haraksim, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic science international 276 (2017) 142-153.

[77] I. Montani, R. Marquis, N. Egli Anthonioz, C. Champod, Resolving differing expert opinions, Science & Justice 59 (2019) 1-8.

[78] A. O'Hagan, C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, T. Rakow, Chapter 4: The Elicitation of Probabilities, Uncertain Judgements: Eliciting Experts' Probabilities, John Wiley & Sons, Ltd, Chichester, UK, 2006.

[79] S. Lichtenstein, B. Fischhoff, L.D. Phillips, Calibration of probabilities: the state of the art to 1980, in: D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases, Cambridge University Press, Cambridge, UK, 1982, pp. 306-334.

[80] L.R. Beach, G.P. Braun, Laboratory studies of subjective probability: A status report, in: G. Wright, P. Ayton (Eds.), Subjective Probability, John Wiley and Sons, Chichester, UK, 1994.

[81] K. Siegel-Jacobs, J.F. Yates, Effects of Procedural and Outcome Accountability on Judgment Quality, Organizational Behavior and Human Decision Processes 65 (1996) 1-17.

[82] J.S. Hammersley, K. Kadous, A.M. Magro, Cognitive and Strategic Components of the Explanation Effect, Organizational Behavior and Human Decision Processes 70 (1997) 149-158.

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

[84] W. Remus, M. O'Conner, K. Griggs, Does Feedback Improve the Accuracy of Recurrent Judgmental Forecasts?, Organizational Behavior and Human Decision Processes 66 (1996) 22-30.

[85] E.R. Stone, R.B. Opel, Training to Improve Calibration and Discrimination: The Effects of Performance and Environmental Feedback, Organizational Behavior and Human Decision Processes 83 (2000) 282-309.

[86] S. Lichtenstein, B. Fischhoff, Training for calibration, Organizational Behavior and Human Performance 26 (1980) 149-171.

[87] G. Keren, Facing uncertainty in the game of bridge: A calibration study, Organizational Behavior and Human Decision Processes 39 (1987) 98-114.

[88] T.S. Wallsten, D.V. Budescu, R. Zwick, S.M. Kemp, Preferences and reasons for communicating probabilistic information in verbal or numerical terms, Bulletin of the Psychonomic Society 31 (1993) 135-138.

[89] I. Erev, B.L. Cohen, Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox, Organizational Behavior and Human Decision Processes 45 (1990) 1-18.

[90] T.S. Wallsten, D.V. Budescu, R. Zwick, Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments, Management Science 39 (1993) 176-190.

[91] E. Reiss, In quest of certainty, The American journal of medicine 77 (1984) 969-971.

[92] G.A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information, Psychological Review 63 (1956) 81-97.

[93] S. Renooij, C.L.M. Witteman, Talking probabilities: communicating probabilistic information with words and numbers, International Journal of Approximate Reasoning 22 (1999) 169-194.

[94] R.M. Hamm, Selection of verbal probabilities: A solution for some problems of verbal probability expression, Organizational Behavior and Human Decision Processes 48 (1991) 193-223.

[95] D.V. Budescu, T.S. Wallsten, Processing Linguistic Probabilities: General Principles and Empirical Evidence, Psychology of Learning and Motivation 32 (1995) 275-318.

[96] T.S. Wallsten, D.V. Budescu, A review of human linguistic probability processing: General principles and empirical evidence, The Knowledge Engineering Review 10 (1995) 43-62.

[97] C.L.M. Witteman, S. Renooij, Evaluation of a verbal–numerical probability scale, International Journal of Approximate Reasoning 33 (2003) 117-131.

[98] K. Fischer, H. Jungermann, Rarely Occurring Headaches and Rarely Occurring Blindness: Is Rarely=Rarely? The Meaning of Verbal Frequentistic Labels in Specific Medical Contexts, Journal of Behavioral Decision Making 9 (1996) 153-172.

Appendix 1

The selected 38 same-source (SS) and 22 different-source (DS) comparisons ranked on the calculated LR based on the 2D data. For each comparison the 2D and 3D comparison score and calculated LRs are given and the difference in degree of support steps between the 2D and 3D LR (Step Δ). In the columns with the header "Correct" it is shown whether the calculated LRs resulted in a true positive (for same-source comparisons) or true negative (for different-source comparisons) result. For the examiners these true positive and true negative results are given as a proportion of their combined judgment.

	Comparison Score		Calculated LR			Correct		
Ground Truth	*2D Data*	*3D Data*	*2D Data*	*3D Data*	*Step Δ*	*2D Data*	*3D Data*	*Examiners (prop)*
SS	0.21	0.32	0	0	0	No	No	0.75
SS	0.47	0.64	0.34	3.38	1	No	Yes	0.45
SS	0.51	0.74	1.71	97.25	1	Yes	Yes	0.53
SS	0.53	0.66	3.11	5.85	0	Yes	Yes	0.75
SS	0.53	0.73	3.52	69.63	1	Yes	Yes	0.96
SS	0.56	0.72	8.55	44.57	1	Yes	Yes	0.91
SS	0.57	0.75	11.69	133.92	1	Yes	Yes	0.92
SS	0.58	0.80	16.97	3765.89	2	Yes	Yes	1.00
SS	0.60	0.65	31.72	4.05	1	Yes	Yes	0.97
SS	0.60	0.77	32.32	422.56	1	Yes	Yes	0.62
SS	0.62	0.62	65.29	1.73	1	Yes	Yes	1.00
SS	0.65	0.85	186.99	17636022.33	3	Yes	Yes	1.00
SS	0.65	0.72	204.92	39.22	1	Yes	Yes	1.00
SS	0.66	0.89	323.65	Infinite	3	Yes	Yes	1.00
SS	0.66	0.67	324.51	10.65	1	Yes	Yes	0.70
SS	0.67	0.68	1043.65	11.49	2	Yes	Yes	0.99
SS	0.68	0.82	1417.36	56249.56	1	Yes	Yes	0.97
SS	0.68	0.90	1459.59	Infinite	2	Yes	Yes	1.00
SS	0.68	0.84	1834.46	782760.68	1	Yes	Yes	0.99

SS	0.69	0.78	2649.85	794.33	1	Yes	Yes	1.00
SS	0.69	0.81	2848.76	8692.25	0	Yes	Yes	0.97
SS	0.70	0.80	2920.77	2767.21	0	Yes	Yes	1.00
SS	0.70	0.89	3013.29	Infinite	2	Yes	Yes	1.00
SS	0.70	0.75	3218.64	151.46	1	Yes	Yes	0.97
SS	0.70	0.94	3802.87	Infinite	2	Yes	Yes	0.99
SS	0.71	0.85	6563.04	19073209.13	2	Yes	Yes	0.99
SS	0.73	0.95	25894.64	Infinite	1	Yes	Yes	1.00
SS	0.73	0.85	53313.82	4659241.26	1	Yes	Yes	0.99
SS	0.73	0.87	88430.28	Infinite	1	Yes	Yes	1.00
SS	0.73	0.90	115953.42	Infinite	1	Yes	Yes	0.99
SS	0.74	0.94	479527.17	Infinite	1	Yes	Yes	1.00
SS	0.88	0.88	Infinite	Infinite	0	Yes	Yes	0.99
SS	0.77	0.89	Infinite	Infinite	0	Yes	Yes	0.99
SS	0.85	0.93	Infinite	Infinite	0	Yes	Yes	1.00
SS	0.90	0.88	Infinite	Infinite	0	Yes	Yes	1.00
SS	0.78	0.89	Infinite	Infinite	0	Yes	Yes	1.00
SS	0.80	0.90	Infinite	Infinite	0	Yes	Yes	1.00
SS	0.83	0.61	Infinite	0.88	6	Yes	No	1.00
DS	0.65	0.58	201.62	0.25	1	No	Yes	0.45
DS	0.63	0.58	103.89	0.25	1	No	Yes	0.60
DS	0.58	0.61	24.04	0.99	1	No	Yes	0.81
DS	0.58	0.63	22.24	2.86	0	No	No	0.90
DS	0.58	0.65	21.52	5.90	0	No	No	0.40
DS	0.57	0.62	16.84	1.73	0	No	No	0.87
DS	0.56	0.58	9.22	0.24	1	No	Yes	0.57
DS	0.53	0.60	3.26	0.80	1	No	Yes	0.69
DS	0.52	0.61	2.88	1.24	0	No	No	0.82
DS	0.51	0.62	1.86	2.04	0	No	No	0.88

DS	0.48	0.60	0.67	0.74	0	Yes	Yes	0.83
DS	0.44	0.59	0.17	0.52	0	Yes	Yes	0.87
DS	0.40	0.56	0.03	0.08	0	Yes	Yes	0.45
DS	0.21	0.40	0.01	0.00	0	Yes	Yes	0.84
DS	0.18	0.40	0.01	0.00	0	Yes	Yes	0.97
DS	0.26	0.26	0.01	0.00	0	Yes	Yes	0.99
DS	0.03	0.41	0.00	0.00	1	Yes	Yes	1.00
DS	0.01	0.42	0	0.00	2	Yes	Yes	0.99
DS	-0.01	0.47	0	0.00	1	Yes	Yes	0.99
DS	-0.02	0.19	0	0.00	1	Yes	Yes	0.97
DS	-0.03	0.44	0	0.00	2	Yes	Yes	0.99
DS	-0.14	0.43	0	0.00	2	Yes	Yes	0.94