Received: 12 May 2020 | Revised: 1 July 2020 | Accepted: 10 August 2020

DOI: 10.1111/1556-4029.14557

JOURNAL OF FORENSIC SCIENCES

PAPER

Criminalistics

Firearm examination: Examiner judgments and computer-based comparisons

Erwin J. A. T. Mattijssen MSc[1,2] | Cilia L. M. Witteman PhD[1] | Charles E. H. Berger PhD[2,3] | Xiaoyu A. Zheng MSc[4] | Johannes A. Soons PhD[4] | Reinoud D. Stoel PhD[2]

[1]Behavioural Science Institute, Radboud University Nijmegen, Nijmegen, The Netherlands

[2]Netherlands Forensic Institute, The Hague, The Netherlands

[3]Institute for Criminal Law and Criminology, Leiden University, Leiden, The Netherlands

[4]Sensor Science Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA

Correspondence
Erwin J.A.T. Mattijssen, Netherlands Forensic Insititute, PO Box 24044, 2490 AA The Hague, The Netherlands.
Email: e.mattijssen@nfi.nl

Present address
Reinoud D. Stoel, Department of Research and Development, Statistics Netherlands, The Hague, The Netherlands

Abstract

Forensic firearm examination provides the court of law with information about the source of fired cartridge cases. We assessed the validity of source decisions of a computer-based method and of 73 firearm examiners who compared breechface and firing pin impressions of 48 comparison sets. We also compared the computer-based method's comparison scores with the examiners' degree-of-support judgments and assessed the validity of the latter. The true-positive rate (sensitivity) and true-negative rate (specificity) of the computer-based method (for the comparison of both the breechface and firing pin impressions) were 94.4% and at least 91.7%, respectively. For the examiners, the true-positive rate was at least 95.3% and the true-negative rate was at least 86.2%. The validity of the source decisions improved when the evaluations of breechface and firing pin impressions were combined and for the examiners also when the perceived difficulty of the comparison decreased. The examiners were reluctant to provide source decisions for "difficult" comparisons even though their source decisions were mostly correct. The correlation between the computer-based method's comparison scores and the examiners' degree-of-support judgments was low for the same-source comparisons to negligible for the different-source comparisons. Combining the outcomes of computer-based methods with the judgments of examiners could increase the validity of firearm examinations. The examiners' numerical degree-of-support judgments for their source decisions were not well-calibrated and showed clear signs of overconfidence. We suggest studying the merits of performance feedback to calibrate these judgments.

KEYWORDS
calibration, comparison algorithm, error rates, expert decision making, forensic firearm examination, inconclusives, judgment, reliability, validity

1 | INTRODUCTION

Forensic firearm examination deals with the analysis and interpretation of features—striations and impressions—in fired cartridge cases and bullets. These features originate from different components of the firearm used to fire the cartridges. Firearm examiners compare these features between cartridge cases or bullets recovered at shooting scenes or between those and reference shots fired with a submitted firearm. They then judge whether the results of this comparison provide support for a same-source or different-source proposition.

In current practice, several reporting formats are applied for these examiner judgments. While there is a growing acceptance that reporting the degree of support for a same-source or different-source proposition is the method of choice [1-3], the uptake of this approach varies between countries and institutes [4]. Categorical conclusions are still often proposed [1], where an examiner reports a decision about the source of a cartridge case or bullet while implicitly assuming some prior odds and cost/benefit assessment of a right/wrong decision. Proficiency tests and most of the literature on the validity and reliability of firearm examiners' judgments are based on the results of examiners providing such decisions. In the remainder of this paper, we will refer to these decisions as "source decisions," while referring to degree-of-support assessments for a source proposition as "degree-of-support judgments." The term "judgments" will be used as an overall label for both types of judgments.

The judgments of the examiners are used in the court of law [5], where they are often treated as impartial and objective [6,7]. Yet, the scientific foundation, and validity and reliability of these judgments, typically provided by human examiners [6,8], are increasingly challenged [5,9] and more empirical research is required on the validity and reliability of the applied methods and resulting judgments [5,9,10]. For the methods, the focus could lie on the implementation of context information management to mitigate unwanted effects of cognitive bias [5,9,11-14]. This could focus on mitigation of bias affecting the observations, their interpretation, and the resulting conclusion [15]. Such methods could include (linear) sequential unmasking [6,16,17], management of case information [18,19], and blind peer review [6,11,20,21]. The validity and reliability of the judgments could be assessed with proficiency tests using items with a known source, where the difficulty of the comparisons is comparable to what is seen in casework. These proficiency tests should preferably be performed in a double-blind fashion in casework [10,22]. Together, these factors could ensure the ecological validity of the proficiency tests [23]. Additionally, calibration studies can be used to assess how well-calibrated the degree-of-support judgments are [24].

Another proposed avenue to improve forensic firearm examination is the development and implementation of more objective, computer-based comparison methods [9]. Currently, the application of such methods is limited to automated database searches, but these methods could also be used to evaluate the evidence. Several such methods have been developed, typically relying on 3D surface topography measurements of striation and impression patterns (e.g., [24,25,26-30]). Computer algorithms compare these topography measurements and provide a comparison score based on the degree of similarity. These comparison scores are then used to make a source decision and to assess the error rate of that decision [27,31], or to determine the evidential strength of a comparison [24,25,30]. The likelihood ratio (LR) is used to express the evidential strength. This LR is the ratio between the probability of the comparison score under two mutually exclusive propositions [32]. An example of a set of such propositions could be as follows: *H1*: The cartridge case was fired with the submitted firearm; and *H2*: The cartridge case was fired with some other firearm. These evaluations require reference data consisting of comparison scores observed for known same-source and known different-source comparisons with characteristics relevant to the case.

With this study, we aim to assess the validity of source decisions made by a computer-based method and by firearm examiners. Furthermore, we will assess the agreement between the examiners' degree-of-support judgments and the outcomes of the computer-based method by looking at the correlation between the two. Additionally, we will assess whether the degree-of-support judgments of the examiners are well-calibrated. We do this to test whether forensic examiners are able to provide accurate degree-of-support judgments based on their training and experience [1,33,34]. For all assessments, we considered the comparison of breechface and firing pin impressions in cartridge cases.

2 | MATERIALS AND METHODS

In this section, we provide information about the test set used in this study, about the computer-based method we used, about the participants and the study design to acquire examiner judgments, and about the analyses we performed.

2.1 | Test material

We created a test with 54 comparison sets. Each comparison set consisted of one "questioned" cartridge case and two reference cartridge cases. The two reference cartridge cases were fired with one firearm. We selected cartridge cases from firearms of different calibers and manufacturers to represent the full range of comparison difficulty seen in casework (as suggested by the American Association for the Advancement of Science (AAAS) [35]). In 38 of the comparison sets, the questioned cartridge case was fired with the same firearm as the two reference cartridge cases (same-source comparisons) and in the other 16 with a different firearm (different-source comparisons). We decided to include more same-source comparisons in the test set because these are more prevalent in casework. For every different-source comparison, the questioned and reference cartridges were fired with firearms with the same caliber, manufacturer, and class characteristics.

We also ensured that the brand of ammunition was the same for all items per comparison set. We randomized the comparison sets into same-source and different-source comparisons and then randomized the order of the comparison sets in the test set. Appendix S1 provides details about the cartridge cases we used and their sources.

2.2 | Computer-based method

2.2.1 | Data acquisition

We measured the 3D topographies of the breechface and firing pin impressions of all cartridge cases in the test set using a disk scanning confocal microscope [36]. For the breechface impressions, we used a 10× magnification objective resulting in a lateral resolution of 3.125 μm and we set the vertical resolution to 10 nm. For the firing pin impressions, we used a 20× magnification objective resulting in a lateral resolution of 1.5625 μm and set the vertical resolution to 5 nm. We manually cropped the measurements to select either the breechface or the firing pin impressions present in the primer cups of the cartridge cases. Doing so, we ensured that only the features of these firearm components were considered during further analysis of the impression patterns. To highlight the firearm features in the topography measurements of the breechface and firing pin impressions, we attenuated the overall shape and measurement noise using Gaussian regression filters with band-pass cutoff lengths (50% attenuation) of 250 μm and 16 μm for the breechface impressions and 200 μm and 8 μm for the firing pin impressions.

To assess the degree of similarity between breechface impressions and firing pin impressions of different cartridge cases, we calculated the normalized Areal Cross Correlation Function maximum ($ACCF_{max}$) [37]. This score represents the average similarity of the compared surface topographies, expressed as the maximum value of the Pearson correlation coefficient observed when changing the relative orientation and position of the compared images. The comparison score ranges from −1 to 1 (maximum negative to maximum positive correlation). An $ACCF_{max}$ of 0 represents no correlation.

2.3 | Examiner judgments

2.3.1 | Participants

We sent an invitation for participation to firearm examiners of forensic institutes and laboratories in North America, South America, Europe, Asia, and Oceania. The letter explained the aims of the study and asked examiners who wanted to participate to sign and return a consent form for the use of their data. Participants could participate individually or as an institute. In the latter case, the comparison sets could be distributed over the examiners within that institute to divide the workload of participation. We informed the participants that they would receive a proficiency report after submitting their answer forms, which we indeed sent them afterward.

Examiners from 32 forensic institutes of all invited continents sent us their answer forms. We received 69 answer forms (see Section 2.3.2) out of the 109 accepted invitations for individual participation and 4 out of the 9 acceptances for participation as an institute. Of the individually participating examiners, 65 stated that they were qualified examiners and 4 that they were in training, and 51 stated that they worked for an accredited institute. Their experience ranged from 1 to 47 years (M = 13.0, Mdn = 12.0, SD = 8.5). Of the participants, 48 stated they normally reported categorical conclusions in casework (e.g., inclusion/inconclusive/exclusion), 16 reported probabilistic conclusions of whom 8 reported likelihood ratios, two reported on a 5-step scale, and one examiner did not provide this information. The examiners who participated as an accredited institute all stated that they were qualified examiners. One of these institutes stated they normally reported categorical conclusions and three reported probabilistic conclusions, one of which as likelihood ratios.

2.3.2 | Study design

Each participating institute received one test set, which could be used by all participating examiners of that institute. The test set consisted of a box with a numbered compartment for each of the 54 comparison sets. For these test sets, we created dark brown pigmented epoxy resin replicas [38] of the original cartridge cases to ensure that all participants examined the same breechface and firing pin impressions. We marked the bottom of each replica with the number of the comparison set and a "Q" or an "R" to identify the questioned and reference cartridge cases, respectively. Furthermore, we marked the location of the extractor mark on the cartridge cases to facilitate positioning of the replicas during examination.

In addition to the test set, the institutes received: (a) an answer form for each participating examiner, (b) information about the design of the test set and instructions how to complete the answer form, with some examples, (c) a short information sheet that summarized the instructions and could be used as a reference during the comparisons, and (d) a list of firearms with which the reference cartridge cases of each comparison set were fired (columns 1–4 of Appendix S1).

Figure 1 shows the layout of the answer form for the first comparison set. We asked the participants to give their judgments for the breechface and firing pin impressions separately. To ensure that the examiners would take into account the same breechface impressions as the computer-based method, we instructed them to only consider the features present in the primer cup. Because the aperture shear marks were often very prominent on the primer cup, we manually scratched these out on the questioned cartridge cases. This ensured that striations of the aperture shear marks could not influence the interpretation of the breechface and firing pin impressions.

As Figure 1 shows, we first asked the participants to indicate if their comparison of the breechface impressions provided support for the proposition that the questioned cartridge case was fired with

Comp Set	Mark to compare	From the same source?	Degree of support						Inconclusive in casework?
1	Breechface	☐ Yes → ☐ NA X ☐ No →	☐ Weak	☐ Moderate	☐ Moderately strong	☐ Strong	☐ Very strong	☐ Extremely strong	☐ Yes ☐ No
	Firing pin	☐ Yes → ☐ NA X ☐ No →	☐ Weak	☐ Moderate	☐ Moderately strong	☐ Strong	☐ Very strong	☐ Extremely strong	☐ Yes ☐ No

FIGURE 1 Layout of the answer form for the first comparison set

the same firearm as the two reference cartridge cases ("Yes") or that it provided support for the proposition that it was fired with a different firearm ("No") in the third column. We refer to these judgments as "source decisions." By using this terminology, we do not mean that an examiner is in a position to make such a categorical decision; there might be other evidence, and it is not up to the examiner to judge the benefit/cost of right/wrong decisions. We stick to the terminology introduced earlier and redefine it as the examiner's opinion about which proposition is supported by the evidence, regardless of how strong that support is. When we refer to the "source decision" of a computer-based method, this means that the resulting likelihood ratio of that method is greater or smaller than one. In practice, our computer-based method did not provide LRs, and we used a proxy instead (see the Section 2.4.2). When the examiners judged that there were no suitable features in the breechface impressions, and they therefore could not make a source decision, they could indicate this by ticking the "not applicable" (NA) box.

Secondly, in the fourth column of Figure 1 we asked the participants to give their judged degree of support for their chosen source proposition on a 6-point scale (Figure 1 and Table 1). Participants could choose to use the verbal scale (e.g., strong support), or the same scale but defined by numerical frequency ranges (e.g., strong support [1 in 1000 to 1 in 10,000 test fires]). We instructed the participants who chose to use the numerical frequency ranges on how to apply these. For same-source decisions, we asked them to assume that the questioned cartridge case was actually fired with a different firearm than the reference cartridge cases, and then to estimate in how many test fires from other firearms (based on the given ranges in Table 1) they would expect to find the same breechface impressions (and resulting degree of similarity). For different-source decisions, we asked them to assume that the questioned cartridge case was actually fired with the same firearm as the reference cartridge cases, and then to estimate in how many test fires from that firearm they would expect to find the same breechface impressions (and resulting low degree or absence of similarity) as in the questioned cartridge case.

We used these estimated frequencies to calculate approximate LRs by considering them as random match equivalents [39] (Table 1). For same-source decisions, we calculated the LR by setting the probability of the comparison results given the same-source proposition to 1 and dividing that by how often the examiners thought they would find the same degree of similarity with test fires from different firearms. For different-source decisions, we set the probability of the comparison results given the different-source proposition to 1. As a result, the LR was determined by how often the examiners thought they would find the same lack of similarity when the questioned cartridge case was fired with the same firearm as the reference cartridges. We decided to apply this approach to approximate LRs because we expected most participants to be fairly unfamiliar with the underlying mechanisms of assigning LRs while this approach still allowed us to study the assumption that examiners can provide accurate degree-of-support judgments. We asked the participant to indicate in the fifth column whether they would be confident to report their source decision in casework or would rather report an "inconclusive" judgment.

TABLE 1 The scale used to indicate the degree of support for the same-source or different-source propositions (column 1) and the corresponding numerical frequency ranges (column 2)

Degree of support	Judged frequency	Approximate LR when support for the same-source proposition was found	Approximate LR when support for the different-source proposition was found
Weak support	In 1 in 2 to 1 in 10 test fires (from different firearms)	2–10	0.1–0.5
Moderate support	In 1 in 10 to 1 in 100 test fires (from different firearms)	10–100	0.01–0.1
Moderately strong support	In 1 in 100 to 1 in 1000 test fires (from different firearms)	100–1000	0.001–0.01
Strong support	In 1 in 1000 to 1 in 10,000 test fires (from different firearms)	1000–10,000	0.0001–0.001
Very strong support	In 1 in 10,000 to 1 in 1,000,000 test fires (from different firearms)	10,000–1,000,000	0.000001–0.0001
Extremely strong support	In less than 1 in 1,000,000 test fires (from different firearms)	>1,000,000	<0.000001

Note: The approximated LRs based on random match equivalents for the same-source and different-source decisions are shown in column 3 and column 4, respectively. The degree-of-support ranges were chosen in line with those published elsewhere [33,54].

After providing these three judgments for the breechface impressions, we asked the participants to do the same for the firing pin impressions. We asked them to repeat these steps for the subsequent 53 comparison sets.

The participants were given approximately four months to perform the comparisons and to return the answer forms. All their answers were manually entered into a spreadsheet, by two people working independently. We then merged the two spreadsheets to check and correct potential mistakes to ensure that all answers were entered correctly.

2.4 | Analyses

2.4.1 | Included comparison sets

After we had sent the test sets to the participants, we discovered that the features in the breechface and the firing pin impressions in comparison sets 6, 16, and 26 provided contradictory source information. We meant these comparison sets to be different-source comparisons, which they were for the breechface impressions but not for the firing pin impressions. As it turned out, the same firing pin was used for the slides of all firearms in the study from which we acquired these cartridge cases [40]. As a consequence, examiners would be expected to find support for the same-source proposition when comparing firing pin impressions, while finding support for the different-source proposition when comparing the breechface impressions of the same comparison set. Because we did not intend such contradictory source information to be present in our test set, we decided not to include these three comparison sets in the analyses.

For comparison sets 14, 48, and 50, we could not use our source decision protocol (see Section 2.4.2) because the caliber and manufacturer of the firearms used to fire the respective cartridge cases were only used once in this study. To mitigate having to create an arbitrary additional source decision proxy, we also removed these three comparison sets, leaving us with 48 comparison sets for the analyses.

2.4.2 | Validity of source decisions

To determine the validity of the computer-based method's and the firearm examiners' source decisions, we calculated the true-positive rates (sensitivity) and true-negative rates (specificity) and the complementary false-negative and false-positive rates. We calculated the true-positive rates (TPR), true-negative rates (TNR), false-positive rates (FPR), and the false-negative rates (FNR) with the following equations: TPR = TP/(TP + FN), TNR = TN/(TN + FP), FPR = FP/(FP + TN), and FNR = FN/(FN + TP). The false-negative rates provide information about the rates of incorrect decisions (reported misleading evidence) for same-source comparisons, and the false-positive rates do so for different-source comparisons. We calculated all four rates for the breechface and firing pin impressions separately and combined. Additionally, we calculated the total error rates of the source decisions, defined as the percentage of incorrect source decisions. We will refer to these five measures (TPR, TNR, FPR, FNR, and total error rate) as the validity measures for the source decisions.

To determine the source decision of the computer-based method, it would seem most appropriate to calculate a likelihood ratio for each comparison (see e.g., [24]). An LR above 1 would provide support for the same-source proposition and an LR below 1 for the different-source proposition. Without considering the degree of support for a proposition, an LR of 1 would then serve as a decision threshold. The same-source and different-source distributions, needed to calculate these LRs, varied between the calibers and firearm manufacturers in our test. Therefore, we judged it to be unadvisable to consider them as one population in this study. This resulted in small subsets of comparison sets per combination of caliber and firearm manufacturer. The number of available cartridge cases per subset was insufficient to calculate robust likelihood ratios. Therefore, we used a proxy to decide whether there is support for the same-source or different-source proposition. This proxy is based on a decision threshold that we aimed to represent an LR value of 1. To calculate this decision threshold, we considered the highest comparison score ($ACCF_{max}$) between the questioned cartridge case and the two reference cartridge cases in each comparison set. We chose to use the highest comparison score because examiners will typically look for the highest degree of similarity with one of the reference cartridge cases. We compared this questioned cartridge case's highest comparison score with the comparison score between the two reference cartridge cases (calculated as the mean of the reversed order comparisons between the two reference cartridge cases, that is, mean of R1 vs R2 and R2 vs R1) in the same comparison set. We also compared it with all different-source comparison scores in the complete test set between cartridge cases that were fired with a firearm of the same caliber and manufacturer. Based on these two results, we defined a source decision protocol to get a proxy for a same-source or different-source decision. Figure 2 shows a schematic representation of this protocol. We defined the midpoint (average) between the comparison score of the two reference cartridge cases and the mean comparison score of the different-source comparisons as a source decision threshold. When the highest comparison score between the questioned cartridge case and the reference cartridge cases was higher than this source decision threshold, we considered this a proxy for a same-source decision. When the highest comparison score was lower than this threshold, we considered this a proxy for a different-source decision. Because of our limited data, we could not consider the dispersion of the same-source comparison scores (which we have no data for) and different-source comparison scores in our source decision protocol. Therefore, we could not approximate the decision threshold (LR = 1) more rigorously than via our considered decision threshold (midpoint between the comparison score of the two reference cartridge cases and the mean comparison score of the different-source comparisons). In an effort to reduce the

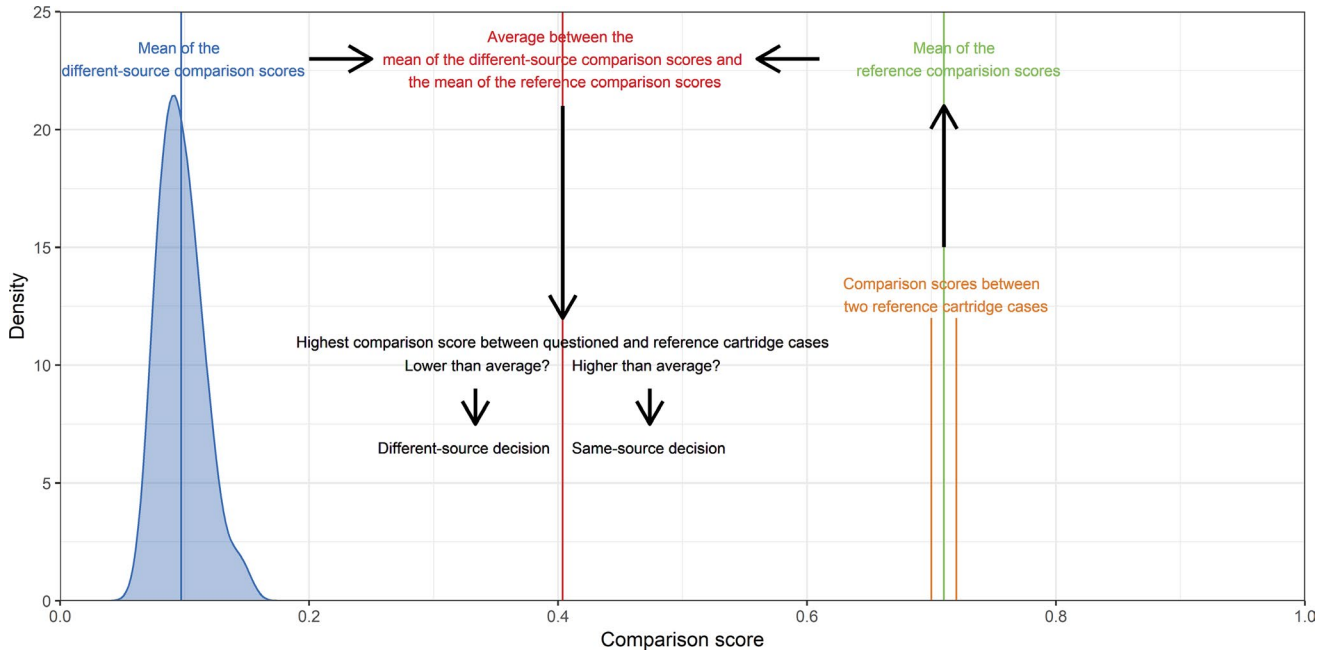

FIGURE 2 Schematic representation of the computer-based method's source decision protocol for a same-source or different-source decision. A different-source comparison score distribution is shown on the left [Color figure can be viewed at wileyonlinelibrary.com]

false-positive rate or the false-negative rate, one could increase or decrease the decision threshold, respectively.

To determine the combined source decisions for the computer-based method, we considered its source decisions of both the breechface and firing pin impressions per comparison set. When the source decisions of these two were the same, we coded the combined source decision similarly. When the source decisions of the two marks were different (one same-source and one different-source decision), we called the combined source decision inconclusive.

To determine the combined source decisions for the examiners, we considered the comparison sets where source decisions for both the breechface and firing pin impression comparisons were given. When the source decisions of these two were the same, we coded the combined source decision similarly. When the source decisions of both marks were different (one same-source and one different-source decision), we considered the magnitude of the two judged degrees of support. When the judged degree of support for the same-source proposition for one of the impression marks was higher than the judged degree of support for the different-source proposition for the other impression mark, the combined source decision would be same source and vice versa. When the degrees of support for contradicting propositions were of the same magnitude, we called the combined source decision inconclusive.

To calculate the validity measures for the source decisions made by the computer-based method and by the examiners, we considered their (by proxy) source decisions. We called a correct same-source decision (for a same-source comparison) a true-positive decision and an incorrect same-source decision (for a different-source comparison) a false-positive decision. Likewise, we called a correct different-source decision (for a different-source comparison) a true-negative decision and an incorrect different-source decision (for a same-source comparison) a false-negative decision.

For the examiners, we calculated the validity measures of all their source decisions and also of only the source decisions that they felt confident to report (excluding the source decisions for which they would have reported an inconclusive judgment in casework). Additionally, we calculated these validity measures for all source decisions per judged degree of support. For the calculation of the validity measures, we did not consider the comparisons where examiners did not report a source decision (missing data) or where they reported a lack of suitable features (indicated by NA in the answer form). We did however compare the examiners' total error rates with their source decision rates. We defined the source decision rate in two ways: (a) as the proportion of the 48 comparisons where they provided a source decision and not an NA judgment or missing data, and (b) as the proportion of the 48 comparisons where they provided a source decision and also felt confident to report this (no NA judgments, missing data, or inconclusive judgment).

We determined the relation between the examiners' degree-of-support judgments and the comparison scores of the computer-based method by calculating the Spearman correlation coefficient. We did this for each examiner, for the same-source and different-source comparisons, and for both the breechface and firing pin impression comparisons separately.

2.4.3 | Calibration of degree-of-support judgments

To determine whether the judgments of the degrees of support were well-calibrated, we looked at the judgments of the examiners who

stated that they had used the numerical frequency ranges as shown in Table 1. The other examiners provided their judgments of the degree of support on a verbal scale, and consequently, we expected lower between-subject reliability in their perception and use of the degrees of support [41,42]. To accommodate peoples' preferences for the use of verbal or numerical scales [43,44] and to aim for the reliable interpretation and use of degree-of-support judgments, it is advised to provide a standardized list of verbal expressions [41,45], that are rank-ordered [46,47], and to define these expressions by (ranges of) numerical probabilities [44,46-50]. Because of the expected low reliability in degree-of-support judgments when the examiners only used the verbal scale and because these judgments were not numerically defined, we do not believe that it would be appropriate to aggregate their judgments to study how well they were calibrated.

If the judgments of the degree of support were well-calibrated, there is a direct link between the judged degree of support and the proportion of incorrect source decisions. The lower the degree of support, the higher the proportion of incorrect source decisions will be and vice versa. We calculated the proportion of incorrect source decisions of the combined examiners and comparisons for each combination of source decision (same source and different source) and degree of support. We compared these proportions with the ranges of expected proportions of incorrect source decisions based on the judged degree of support. We calculated these expected ranges based on the approximated LR ranges (Table 1), using Equation (1) [51]:

$$\frac{1}{\frac{N_{\text{same-source comparisons}}}{N_{\text{different-source comparisons}}}\times \text{LR}+1}=\frac{1}{\frac{36}{12}\times \text{LR}+1} \quad (1)$$

This equation is based on the premises that an ideally calibrated degree-of-support judgment correctly represents the evidential strength (LR) of the findings of the comparison. When this is true, the same probabilities that apply to the findings of the comparison should also apply to the LR: That is, the probability of the findings and of the corresponding LR should both be *x* times larger given *H1* than given *H2*. So, when the findings (*E*) are—for example—10 times more probable under *H1* than under *H2* (LR =10), the LR of 10 should also be 10 times more probable under *H1* than under *H2* (Equation 2).

$$\frac{P(E|H1)}{P(E|H2)}=\text{LR}=\frac{P(\text{LR}|H1)}{P(\text{LR}|H2)} \quad (2)$$

Assuming this ideal calibration and an equal number of same-source and different-source comparisons (a priori ratio of 1), Equation (1) is derived from Equation (3) [51].

$$\frac{P(\text{LR}|H2)}{P(\text{LR}|H1)+P(\text{LR}|H2)}=\frac{1}{P(\text{LR}|H1)/P(\text{LR}|H2)+1}=\frac{1}{\text{LR}+1} \quad (3)$$

When the number of same-source ($N_{\text{same-source comparisons}}$) and different-source comparisons ($N_{\text{different-source comparisons}}$) is not equal, the LR value in Equation (3) should be multiplied by $N_{\text{same-source comparisons}}/N_{\text{different-source comparisons}}$, leading to Equation (1).

We considered the judged degree of support to be well-calibrated when the actual proportion of incorrect source decisions fell within the range of expected proportions of incorrect source decisions. We also calculated the Pearson correlation coefficient between the actual proportions of incorrect source decisions and the upper bound of the range of expected proportions of incorrect source decisions. We assessed the calibration of degree-of-support judgments and their correlation with the range of expected proportions of incorrect source decisions for the same-source and different-source decisions and for the breechface and firing pin impressions separately.

3 | RESULTS

3.1 | Validity of source decisions

The validity measures of the source decisions for the comparisons of the breechface and firing pin impressions are shown in the confusion matrices in Table 2. These are shown for the computer-based method, all examiners' source decisions, and the examiners' source decisions that they felt confident to report.

When we explore the validity measures in Table 2, we see that the true-positive rates of the examiners for both the breechface and firing pin impression comparisons are slightly higher than those of the computer-based method's source decisions (Table 2), while the true-negative rates are slightly higher for the computer-based method. When we only consider the examiners' source decisions that they felt confident to report, we see that their true-positive and true-negative rates are slightly higher than when we consider all their source decisions. But even then, the examiners' true-negative rates are not as high as those of the computer-based method. The total error rates of the examiners' source decisions are also higher than those of the computer-based method.

When we look at how often examiners did not feel confident to report their observed support for a same-source or different-source proposition, we see that the majority of the examiners' inconclusive judgments are given for initial true-positive and true-negative source decisions, both for the breechface (80.3%) and for the firing pin (80.7%) impression comparisons.

While Table 2 shows the validity measures for the examiners as a group, there are large individual differences (Figures S1 and S2). Not only did these validity measures vary between examiners, the examiners also differed in how often they decided that there were no suitable features in the impressions to provide support for a same-source or different-source proposition (NA) and in how often they did not feel confident to report their source decision (inconclusive judgment). Figure 3 shows the relation between the examiners' total error rates and their source decision rates.

There was an overall positive relation between validity of source decision and degree of support (Figure 4 and Appendix

TABLE 2 Confusion matrices and total error rates for the same-source (SS) and different-source (DS) decisions of the computer-based method and the examiners

Breechface impression comparisons				**Firing pin impression comparisons**			
Computer-based method				*Computer-based method*			
N = 48	SS comparison	DS comparison		*N* = 48	SS comparison	DS comparison	
SS decision	34	1	FPR = 0.083	SS decision	34	0	FPR = 0.000
DS decision	2	11	FNR = 0.056	DS decision	2	12	FNR = 0.056
	TPR = 0.944	TNR = 0.917			TPR = 0.944	TNR = 1.000	
Total error rate	6.25%			Total error rate	4.17%		
All examiners' source decisions				*All examiners' source decisions*			
N = 3504	SS comparison	DS comparison		*N* = 3504	SS comparison	DS comparison	
SS decision	2009	101	FPR = 0.138	SS decision	2287	98	FPR = 0.127
DS decision	97	632	FNR = 0.046	DS decision	114	676	FNR = 0.047
	TPR = 0.954	TNR = 0.862			TPR = 0.953	TNR = 0.873	
Total error rate	6.97%			Total error rate	6.68%		
No features to compare	597			No features to compare	245		
Missing source decisions	68			Missing source decisions	84		
Examiners' source decisions that they felt confident to report				*Examiners' source decisions that they felt confident to report*			
N = 2767	SS comparison	DS comparison		*N* = 2705	SS comparison	DS comparison	
SS decision	1635	52	FPR = 0.112	SS decision	1913	56	FPR = 0.121
DS decision	48	414	FNR = 0.029	DS decision	55	405	FNR = 0.028
	TPR = 0.971	TNR = 0.888			TPR = 0.972	TNR = 0.879	
Total error rate	4.65%			Total error rate	4.57%		
No features to compare	597			No features to compare	245		
Missing source decisions	21			Missing source decisions	31		
Number of inconclusive judgments				*Number of inconclusive judgments*			
N = 737	SS comparison	DS comparison	Total	*N* = 799	SS comparison	DS comparison	Total
SS decision	374	49	423	SS decision	374	42	416
DS decision	49	218	267	DS decision	59	271	330
Total	423	267			433	313	
Missing source decisions	47			Missing source decisions	53		

Note: This table also shows how often the examiners judged that there were no features to compare (NA), the number of missing source decisions (where no source decision was made on the answer form), and the number of inconclusive judgments.

S2). The higher the examiners' judged the degree of support for a source proposition, the higher the true-positive and true-negative rates were and the lower the complementary false-negative and false-positive rates, as well as the total error rate.

The validity measures in Table 3 show that the computer-based method's source decisions for the breechface and firing pin impression comparisons combined are comparable to or better than those for the separate impression mark types. The total error rate of the combined source decisions is lower than that of the separate source decisions (Table 2). The examiners' source decisions combined (Table 3) are also better than those of their source decisions for these impression marks separately, and the total error rate of these source decisions combined is also lower than that of the separate impression marks (Table 2).

3.2 | Relation between examiners' degree-of-support judgments and the outcomes of the computer-based method

We found low positive Spearman correlations for the same-source comparisons for both the comparison of breechface (M = 0.38, SD = 0.19, 95% CI [0.34, 0.42]) and firing pin impressions (M = 0.32, SD = 0.20, 95% CI [0.27, 0.36]) [52]. For the different-source

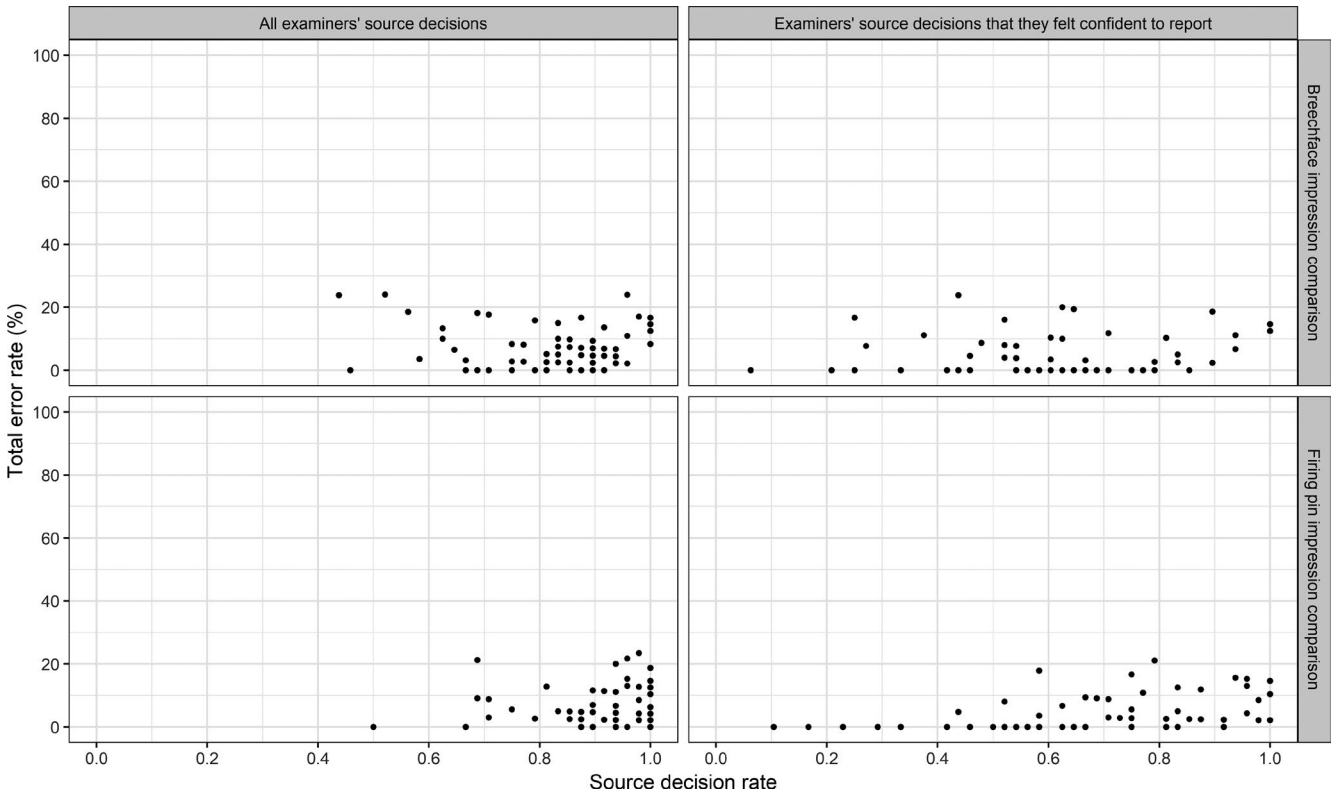

FIGURE 3 The total error rate as a function of the source decision rate per examiner for the breechface impression (top) and firing pin impression (bottom) comparisons. In the left two panels, the source decision rate is calculated as the proportion of the 48 comparisons where the examiner provided a source decision and not an NA judgment or missing data. In the right two panels, the source decision is calculated as the proportion of the 48 comparisons where the examiner provided a source decision and also felt confident to report this (no NA judgments, missing data, or inconclusive judgments)

comparisons, we found negligible positive correlations for the comparison of breechface impressions (M = 0.11, SD = 0.33, 95% CI [0.03, 0.19]) and negligible negative correlations for the firing pin impressions (M = −0.07, SD = 0.35, 95% CI [−0.15, 0.01]).

3.3 | Calibration of degree-of-support judgments

The examiners of 21 of the 73 answer forms stated that they had used the numerical ranges provided (Table 2) when judging the degrees of support for their source decisions. Figure 5 shows the relation between the proportion of their incorrect source decisions and their judged degrees of support for both the breechface and firing pin impression comparisons. In this figure, we also show the ranges of expected proportions of incorrect source decisions given the judged degree of support (shaded area).

For the breechface impression comparisons, we found a significant and very high positive Pearson correlation coefficient between the upper bound of the range of expected proportions of incorrect source decisions and the proportion of actual incorrect same-source decisions ($r(4)$ =0.92, p = 0.008) and no significant correlation for the different-source decisions (p = 0.238). For the firing pin impression comparisons, we found a significant and very high positive Pearson correlation coefficient for the same-source decisions ($r(4)$ =0.97, p = 0.002) and no significant correlation for the different-source decisions (p = 0.834). However, even when we observe a high positive correlation, most of the actual proportions of incorrect source decisions are a lot higher than would be expected based on the examiner's judged degrees of support (Figure 5).

4 | DISCUSSION

This study showed the performance of the computer-based method and firearm examiners when giving support for propositions about the source of a questioned cartridge case based on the comparison of breechface or firing pin impressions. The true-positive rates (sensitivity) and true-negative rates (specificity) for the comparisons of these impressions by either the computer-based method or the examiners as a group ranged from 86.2% to 95.4%. When we compare the examiners' validity measures with those of the computer-based method, we see that the examiners performed slightly better at identifying same-source comparisons correctly and the computer-based method performed slightly better at identifying different-source comparisons correctly. Of course, these results do not only depend on performance but also on the chosen decision thresholds. This result is not consistent with the results of earlier studies [24,53], but the differences between the computer-based

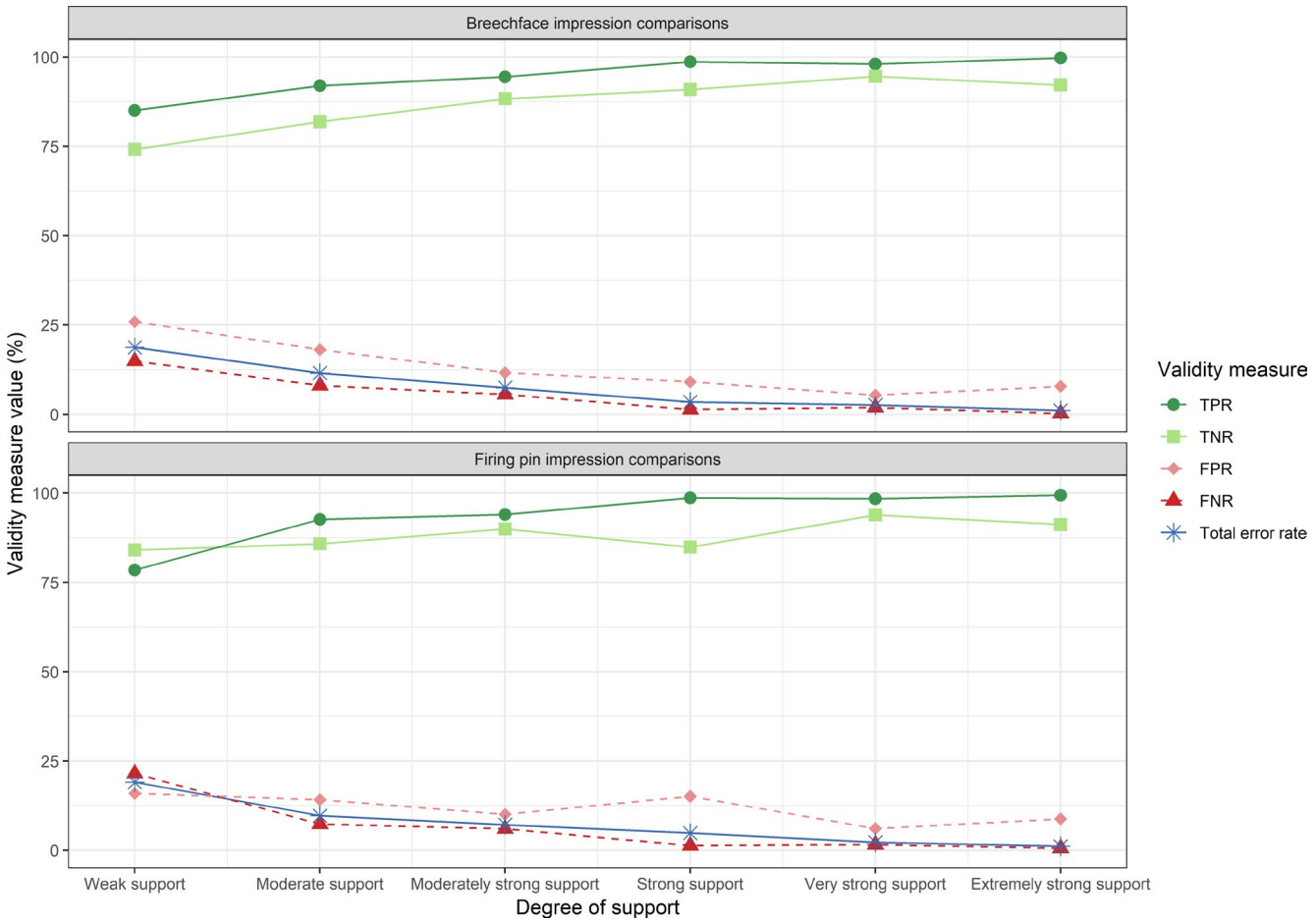

FIGURE 4 Values for each of the source decision validity measures based on all examiners' source decisions per judged degree of support for both the breechface and firing pin impression comparisons [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Confusion matrices and total error rates for the combined source decisions of both the computer-based method and examiners

Computer-based method			
N = 48	SS comparison	DS comparison	
SS decision	33	0	FPR = 0.000
DS decision	1	11	FNR = 0.029
	TPR = 0.971	TNR = 1.000	
Total error rate	2.22%		
Inconclusive results	3		
All examiners' source decisions			
N = 3504	SS comparison	DS comparison	
SS decision	1921	79	FPR = 0.118
DS decision	55	591	FNR = 0.028
	TPR = 0.972	TNR = 0.882	
Total error rate	5.06%		
No combined source decision	858		

Note: For the computer-based method, the number of inconclusive results is given, and for the examiners, the number of comparisons where one of the source decisions was missing.

method's and examiners' performance are small, and we used a different computer-based method than used in the earlier studies. The validity measures of the examiners' source decisions improved when the examiners were given the liberty to decide which of their source decisions they felt confident to report. This outcome corresponds to the results of a study by Mattijssen et al. [24]. But even when we only consider the source decisions that they felt confident to report, the examiners' true-negative rates were not as high as those

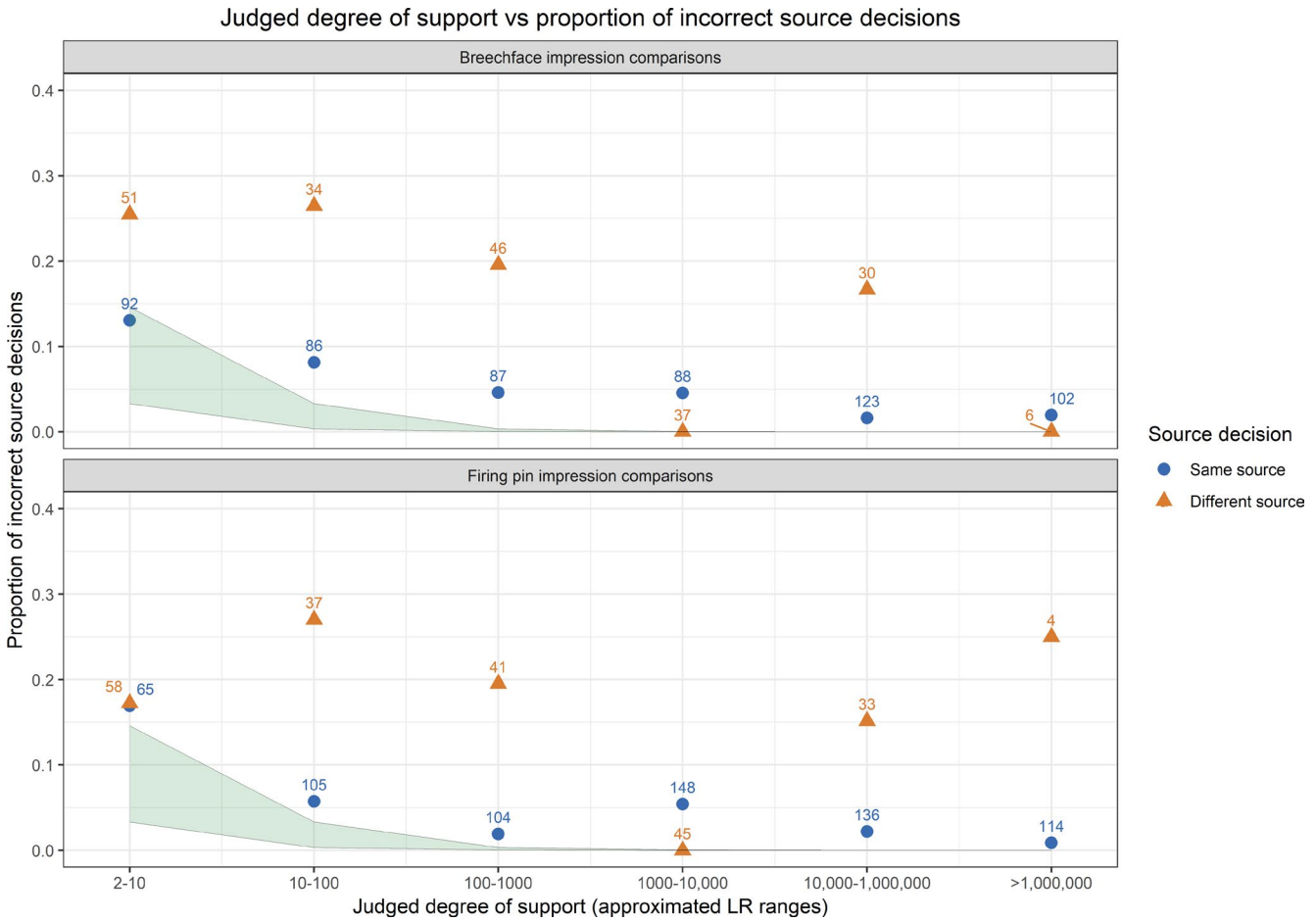

FIGURE 5 The proportion of incorrect source decisions per judged degree of support for both the same-source and different-source decisions and the number of source decisions per combination. For the different-source decisions, we reversed the source propositions to calculate the approximate LR ranges (1/LR). The shaded area represents the expected range of incorrect source decisions given the judged degree of support [Color figure can be viewed at wileyonlinelibrary.com]

of the computer-based method. We believe that we should focus on all examiners' source decisions for a fair comparison between the computer-based method and the examiners, because considering all source decisions does not enable the examiners to artificially alter the overall difficulty of the test set. By stating that they did not feel confident to report their source decisions, examiners could remove "difficult" comparisons from the test set. This is a liberty that the computer-based method did not have. The difference in the possibility to affect the difficulty of the test between the examiners and the computer-based method was not completely mitigated, however, as the validity measures of the examiners' source decisions did not consider comparisons where examiners did not report a source decision (missing data) or where they reported a lack of suitable features for comparison (NA judgments). The computer-based method was set to provide a source decision for all comparison sets, which may have increased its error rates relative to those of the examiners.

Figure 3 shows that there were large differences between examiners in how often they stated that there were no suitable features for comparisons or that they did not feel confident to report their judged support for a same-source or different-source proposition (inconclusive judgement). The inclination to state that there are no suitable features for comparison varied considerably between examiners, with examiners' source decision rates ranging from 0.44 to 1. When the examiners were also allowed to state that they did not feel confident to report their source decision (inconclusive judgments), their source decision rates ranged from 0.06 to 1. These differences are striking, especially when we consider that the use of the epoxy resin replicas ensured that all examiners were provided with the same features for comparison. The differences in the inclination of examiners to provide source decisions show that it may be hard to predict the "appropriate" outcome of a given comparison when setting up a proficiency test where examiners are asked to provide a categorical conclusion about the source of a cartridge case [23]. Asking examiners to judge the degree of support for a source proposition will enable them to provide more information, without having to adhere to decision thresholds, which apparently vary considerably between examiners.

When we look at the examiners' total error rates, we also see considerable differences, with total error rates ranging from 0% to 24%. The observed variability in both the source decision rates and the total error rates indicates that there is room to improve the reliability of examiner judgments. It can be expected that examiners

who are reluctant to provide source decisions (high decision threshold) would provide fewer incorrect source decisions, but this does not always seem to be the case. These results show that trying to establish one overall error rate for a forensic discipline is counter-productive [23,24]. The error rate (or rate of misleading evidence) will depend on both the difficulty of the comparison and the examiner's expertise. It could be argued that an examiner's expertise could be assessed by considering both their total error rate and their source decision rate per perceived degree of difficulty of comparison. Performance would increase with decreasing error rates and increasing source decision rates. Individual or group results over time could be used to assess, for example, the effects of additional training or of new quality control procedures [23].

We consider the judged degrees of support for the source propositions as a proxy for perceived comparison difficulty; the lower the judged degree of support, the more difficult the comparison. The results that the true-positive and true-negative rates increased and that the total error rates decreased with decreasing difficulty (increasing degree of support) indicate that the validity of source decisions depends on the comparison difficulty. These results correspond to those of another study on the validity of forensic firearm examiners' source decisions [24]. Based on these results, we recommend the use of a reporting format in which the evidential strength of a comparison is reported in probabilistic terms (a likelihood ratio) to provide the court of law with a more informative opinion. This recommendation is in line with that of others [4,34,35,54,55] and corresponds to the growing acceptance of the logical approach as the method of choice for interpreting forensic evidence [1-3].

Combining the separate source decisions for the breechface and firing pin impression comparisons had an overall positive effect on the performance of both the computer-based method and the examiners. This is clearly seen when looking at the total error rates, which decrease from 6.25% (breechface impressions) and 4.17% (firing pin impressions) to 2.22% (combined) for the computer-based method and from 6.97% (breechface impressions) and 6.68% (firing pin impressions) to 5.06% for the examiners. These results are comparable to those of an earlier study [56] and suggest that when possible it is beneficial to combine the features of several firearm marks when judging the source of a cartridge case instead of basing an opinion on the features of just one mark (assuming that no firearm components have been replaced).

The higher validity measures, which were observed when the information from the breechface and firing pin impressions is combined or when examiners judged the degree of support for a proposition to be higher, seem comparable to the performance of examiners in earlier studies (e.g., [24,57,58]).

We acknowledge that it is difficult to relate the computer-based method's and the examiners' validity measures from this study to their performance in actual casework (see also [23,59]). We selected cartridge cases from various calibers and firearm manufacturers to represent the full range of comparison difficulty seen in casework, but are unsure whether the resulting test set really provided a representative sample (equally difficult) of the relevant population, or a less or more difficult sample. Furthermore, we asked the examiners to judge the source of cartridge cases based on the comparison of two separate impression marks. In actual casework, examiners can combine the results of these judgments with those of other firearm marks, which this study indicates to be beneficial for the validity of the source decisions. The use of epoxy resin replicas could also have affected the outcomes. We chose to use these replicas to ensure that all examiners examined the same breechface and firing pin impressions. A downside of using replicas is that visualizing the features in the breechface and firing pin impressions is typically somewhat more difficult than in original cartridge cases. This could have increased the difficulty of the comparisons. We tried to minimize the negative effect of using replicas by providing the examiners with the possibility to state that there were no suitable features to compare (NA judgment). Also, the examiners' source decisions were not peer reviewed, while this is typically done in actual casework with the aim to improve the validity of source decisions and to prevent errors [13,20,21,60]. The examiners might also have behaved differently because they knew they were participating in a study in which the actual sources of cartridge cases were known and would receive performance feedback, which is not possible in actual casework. Finally, this study did not include some forms of context information that are typically seen in casework. In casework, bias resulting from—for example—the case information, and cultural and organizational factors could also affect both the observations of the examiners and the subsequent interpretation of those observations [61-63].

The result that approximately 80% of the inconclusive judgments were given when the initial source decision (either supporting the same-source or different-source proposition) was actually correct indicates that examiners seemed reluctant to report source decisions. In an earlier study, a similar result was seen, where approximately 73% of the inconclusive judgments were given when the source decisions were correct [24]. We believe that these results show that examiners, who are expected to provide impartial and objective judgments [6,7], are reluctant to provide a source decision, which could potentially be incorrect. They seem to prefer to default to the current noninformative state, which could be perceived as less harmful to the judicial decision-making process than a wrong opinion with the weight of a categorical conclusion. The fact that most participants (67%) typically provide categorical conclusions in casework could have contributed to this finding. Such a source decision implies a high level of certainty, and examiners could have applied a high decision threshold to guard against providing an incorrect source decision, defaulting to an inconclusive judgment. If examiners would refrain from expressing categorical conclusions and would express their judged degree of support for a source proposition, they would be able to provide the court of law with valuable information in more cases. When expressing the degree of support, inconclusive judgments (or “approximately equally probable” LR conclusions) of forensic examiners should be equally distributed over correct and incorrect source decisions (see [64] for a discussion about the use

of inconclusive judgments). In this study, this was not the case. An inconclusive judgment was 7.6 (for breechface impression comparisons) or 8.9 (for firing pin impression comparisons) times more likely when the initial same-source decision was correct than when it was incorrect. A correct initial different-source decision was 4.4 (breechface impression comparisons) or 4.6 (firing pin impression comparisons) times more likely to result in an inconclusive judgment than an incorrect initial different-source decision. These results seem to indicate that for difficult comparisons, examiners were more certain about their incorrect source decision than about their correct source decision, while also being slightly more reluctant to report a same-source than a different-source decision.

The low positive correlations between the examiners' degree-of-support judgments and the comparison scores of the computer-based method for the same-source comparisons do not provide strong support for their agreement. The negligible associations for the different-source comparisons provide no support for agreement. These low to negligible associations could in part be explained by possible differences in comparison approaches used by the computer-based method and the examiners. The computer-based method's comparison score represents the average similarity of the compared impressions, while examiners have the liberty to focus on specific features instead of on the whole surface area of these impressions, thus considering different information. Because of such possible differences in comparison approaches, future work could explore possibilities of combining the outcomes of computer-based methods with the judgments of examiners to increase the overall validity of firearm examination.

When we compare the correlation results with those of another study [24], we see that those of the same-source comparisons are comparable, while those of the different-source comparisons are lower in this study (from $r \approx 0.5$ to $r \approx 0.1$). We are unsure what caused this deviation for the different-source comparisons, but it could be related to the differences in study design. In the study of Mattijssen et al. [24], only the firing pin aperture shear marks of cartridge cases fired with 9 mm Luger Glock pistols were compared, while in this study, the features of other marks in cartridge cases of various calibers and firearm manufacturers were compared. The added variability in features and the chance that some of the examiners might be less familiar with some of these firearm manufacturers could have resulted in more variability in the degree-of-support judgments, decreasing the potential for a positive association with the computer-based method's comparison scores. Due to the low and negligible associations between the computer-based method and the examiners, it is likely that their reported degree-of-support judgments for the same comparison will deviate. Such differing outcomes could be resolved in practice [65], where the combination of the outcomes could increase the overall validity of firearm examination.

On 21 out of the 73 answer forms, the examiners stated that they had used the numerical ranges provided to judge the degree of support for a source proposition. When we look at how well these degree-of-support judgments are calibrated (Figure 5), we see that for most degrees of support, the actual proportions of incorrect source decisions are a lot higher than would be expected. This is especially evident for the different-source decisions. The fact that most of the actual proportions of incorrect source decisions are too high means that the judged degrees of support were too high. These results are similar to those of an earlier study that looked at how well the degree-of-support judgments of firearm examiners were calibrated [24]. The study of Mattijssen et al. [24] and the current study indicate overconfidence of examiners, where their estimates for the degree of support for a same-source or different-source proposition are too high. This overconfidence is also seen in other studies on the calibration of judgments of expert groups [66-69]. It does not support the assumption that examiners are capable of providing accurate degree-of-support judgments based on their training and experience [1,33,34]. Since the court of law relies on these judgments, it is important that they are well-calibrated. Performance feedback has been shown to increase calibration of judgments [66,70-72], and the merits of this technique should be studied for the calibration of examiner judgments.

5 | CONCLUSIONS

We conclude that the true-positive rate (sensitivity) and true-negative rate (specificity) of the computer-based method (for the comparison of both the breechface and firing pin impressions) were 94.4% and at least 91.7%, and that for the examiners, the true-positive rate was at least 95.3% and the true-negative rate was at least 86.2%. The validity of the source decisions improved when the information from different impression marks was combined. The validity of the examiners' source decisions also improved when the comparison was perceived as less difficult, as assessed by the estimated degree of support for a source proposition. The correlation between the comparison scores of the computer-based method and the degree-of-support judgments of the examiners was low for the same-source comparisons and negligible for the different-source comparisons. The examiners' judged numerical degrees of support for the source decisions were not well-calibrated and showed clear signs of overconfidence: They were too high when compared to the respective proportions of incorrect source decisions. Examiners were reluctant to provide initial source decisions for "difficult" comparisons, even though in this study, the majority of these initial source decisions were correct. This demonstrates an important aspect of reporting categorical conclusions: It does not allow one to report evidence that provides valuable information but less than certainty.

ACKNOWLEDGEMENTS

The authors would like to thank all examiners who participated in this study. Furthermore, we would like to thank the Prince George's County Police Department (PGPD), FBI, and ATF for firing some of the cartridge cases we used, and the authors of previous studies (55 70,93-95) for providing the National Institute of Standards and Technology with cartridge cases of their studies to be used for other

research purposes. The authors would also like to thank Michael Stocker, Thomas Renegar, and Derrel Garcia for their help in creating the replicas of the cartridge cases used in this study.

REFERENCES

1. Biedermann A, Garbolino P, Taroni F. The subjectivist interpretation of probability and the problem of individualisation in forensic science. Sci Justice. 2013;53(2):192–200. https://doi.org/10.1016/j.scijus.2013.01.003.
2. Aitken CGG, Berger CEH, Buckleton JS, Champod C, Curran J, Dawid AP, et al. Expressing evaluative opinions: A position statement. Sci Justice. 2011;51(1):1–2. https://doi.org/10.1016/j.scijus.2011.01.002.
3. Martire KA, Kemp RI, Sayle M, Newell BR. On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. Forensic Sci Int. 2014;240:61–8. https://doi.org/10.1016/j.forsciint.2014.04.005.
4. Martire KA, Kemp RI, Watkins I, Sayle MA, Newell BR. The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. Law Hum Behav. 2013;37(3):197–207. https://doi.org/10.1037/lhb0000027.
5. Committee on Identifying the Needs of the Forensic Sciences Community: National Research Council. Strengthening forensic science in the United States: a path forward. Washington, DC: The National Academies Press, 2009.
6. Kassin SM, Dror IE, Kukucka J. The forensic confirmation bias: problems, perspectives, and proposed solutions. J Appl Res Mem Cogn. 2013;2(1):42–52. https://doi.org/10.1016/j.jarmac.2013.01.001.
7. Dror IE, Kassin SM, Kukucka J. New application of psychology to law: Improving forensic evidence and expert witness contributions. J Appl Res Mem Cogn. 2013;2(1):78–81. https://doi.org/10.1016/j.jarmac.2013.02.003.
8. Dror IE, Cole SA. The vision in "blind" justice: expert perception, judgment, and visual cognition in forensic pattern recognition. Psychon Bull Rev. 2010;17(2):161–7. https://doi.org/10.3758/pbr.17.2.161.
9. Executive Office of the President's Council of Advisors on Science and Technology. Report to the President—Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf. Accessed 10 Aug 2020.
10. Mnookin JL, Cole SA, Dror IE, Fisher BAJ, Houck MM, Inman K, et al. The need for a research culture in the forensic sciences. UCLA Law Rev. 2011;58:725–79. https://doi.org/10.2139/ssrn.1755722.
11. Risinger DM, Saks MJ, Thompson WC, Rosenthal R. The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. Calif Law Rev. 2002;90(1):1–56. https://doi.org/10.2307/3481305.
12. Cole SA. Implementing counter-measures against confirmation bias in forensic science. J Appl Res Mem Cogn. 2013;2(1):61–2. https://doi.org/10.1016/j.jarmac.2013.01.011.
13. Dror IE. Practical solutions to cognitive and human factor challenges in forensic science. Forensic Sci Pol Manag: Int J. 2013;4(3–4):105–13. https://doi.org/10.1080/19409044.2014.901437.
14. Cooper GS, Meterko V. Cognitive bias research in forensic science: A systematic review. Forensic Sci Int. 2019;297:35–46. https://doi.org/10.1016/j.forsciint.2019.01.016.
15. Dror IE. A hierarchy of expert performance. J Appl Res Mem Cogn. 2016;5(2):121–7. https://doi.org/10.1016/j.jarmac.2016.03.001.
16. Krane DE, Ford S, Gilder JR, Inman K, Jamieson A, Koppl R, et al. Sequential unmasking: A means of minimizing observer effects in forensic DNA interpretation. J Forensic Sci. 2008;53(4):1006–7. https://doi.org/10.1111/j.1556-4029.2008.00787.x.
17. Dror IE, Thompson WC, Meissner CA, Kornfield IL, Krane DE, Saks MJ, et al. Letter to the Editor - Context management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. J Forensic Sci. 2015;60(4):1111–2. https://doi.org/10.1111/1556-4029.12805.
18. Mattijssen EJAT, Kerkhoff W, Berger CEH, Dror IE, Stoel RD. Implementing context information management in forensic casework: Minimizing contextual bias in firearms examination. Sci Justice. 2016;56(2):113–22. https://doi.org/10.1016/j.scijus.2015.11.004.
19. Found B, Ganas J. The management of domain irrelevant context information in forensic handwriting examination casework. Sci Justice. 2013;53(2):154–8. https://doi.org/10.1016/j.scijus.2012.10.004.
20. Ballantyne KN, Edmond G, Found B. Peer review in forensic science. Forensic Sci Int. 2017;277:66–76. https://doi.org/10.1016/j.forsciint.2017.05.020.
21. Mattijssen EJAT, Witteman CLM, Berger CEH, Stoel RD. Cognitive biases in the peer review of bullet and cartridge case comparison casework: A field study. Sci Justice. 2020;60(4):337–46. https://doi.org/10.1016/j.scijus.2020.01.005.
22. Kerkhoff W, Stoel RD, Berger CEH, Mattijssen EJAT, Hermsen R, Smits N, et al. Design and results of an exploratory double blind testing program in firearms examination. Sci Justice. 2015;55(6):514–9. https://doi.org/10.1016/j.scijus.2015.06.007.
23. Dror IE. The error in "error rate": Why error rates are so needed, yet so elusive. J Forensic Sci. 2020;65(4):1034–9. https://doi.org/10.1111/1556-4029.14435.
24. Mattijssen EJAT, Witteman CLM, Berger CEH, Brand NW, Stoel RD. Validity and reliability of forensic firearm examiners. Forensic Sci Int. 2020;307:110112. https://doi.org/10.1016/j.forsciint.2019.110112.
25. Riva F, Champod C. Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases. J Forensic Sci. 2014;59(3):637–47. https://doi.org/10.1111/1556-4029.12382.
26. Zheng X, Soons J, Vorburger TV, Song J, Renegar T, Thompson R. Applications of surface metrology in firearm identification. Surf Topogr Metrol. 2014;2(1):014012. https://doi.org/10.1088/2051-672x/2/1/014012.
27. Song J. Proposed, "congruent matching cells (CMC)" method for ballistic identification and error rate estimation. AFTE J. 2015;47:177–85.
28. Gambino C, McLaughlin P, Kuo L, Kammerman F, Shenkin P, Diaczuk P, et al. Forensic surface metrology: Tool mark evidence. Scanning. 2011;33(5):272–8. https://doi.org/10.1002/sca.20251.
29. Baiker M, Keereweer I, Pieterman R, Vermeij E, van der Weerd J, Zoon P. Quantitative comparison of striated toolmarks. Forensic Sci Int. 2014;242:186–99. https://doi.org/10.1016/j.forsciint.2014.06.038.
30. Riva F, Mattijssen EJAT, Hermsen R, Pieper P, Kerkhoff W, Champod C. Comparison and interpretation of impressed marks left by a firearm on cartridge cases – Towards an operational implementation of a likelihood ratio based technique. Forensic Sci Int. 2020;313:110363. https://doi.org/10.1016/j.forsciint.2020.110363.
31. Petraco NDK, Kuo L, Chan H, Phelps E, Gambino C, McLaughlin P, et al. Estimates of striation pattern identification error rates by algorithmic methods. AFTE J. 2013;45:235–44.
32. Aitken CGG, Taroni F. Statistics and the evaluation of evidence for forensic scientists, 2nd edn. Chichester, UK: John Wiley and Sons, 2004;95–9.
33. European Network of Forensic Science Institutes. ENFSI guideline for evaluative reporting in forensic science – Strengthening

the evaluation of forensic results across Europe (STEOFRAE). Wiesbaden, Germany: ENFSI, 2016; Report No.: HOME/2010/ISEC/MO/4000001759.
34. Nordgaard A, Ansell R, Drotz W, Jaeger L. Scale of conclusions for the value of evidence. Law Probab Risk. 2012;11(1):1–24. https://doi.org/10.1093/lpr/mgr020.
35. AAAS. Forensic science assessments: A quality and gap analysis – latent fingerprint examination. Washington, DC: The American Association for the Advancement of Science, 2017. https://doi.org/10.1126/srhrl.aag2874.
36. Hamilton DK, Wilson T. Three-dimensional surface measurement using the confocal scanning microscope. Appl Phys B. 1982;27(4):211–3. https://doi.org/10.1007/BF00697444.
37. Vorburger TV, Song J, Petraco N. Topography measurements and applications in ballistics and tool mark identifications. Surf Topogr Metrol. 2016;4(1):013002. https://doi.org/10.1088/2051-672x/4/1/013002.
38. Koch A. Castings of complex stereometric samples for proficiency tests in firearm and tool mark examinations. AFTE J. 2007;39(4):299–306.
39. Thompson WC. Discussion paper: Hard cases make bad law—reactions to R v T. Law Probab Risk. 2012;11(4):347–59. https://doi.org/10.1093/lpr/mgs020.
40. LaPorte D. An empirical and validation study of breechface marks on.380 ACP caliber cartridge cases fired from ten consecutively finished hi-point model C9 pistols. AFTE J. 2011;43(4):303–9.
41. Budescu DV, Wallsten TS. Consistency in interpretation of probabilistic phrases. Organ Behav Hum Dec. 1985;36(3):391–405. https://doi.org/10.1016/0749-5978(85)90007-X
42. Budescu DV, Weinberg S, Wallsten TS. Decisions based on numerically and verbally expressed uncertainties. J Exp Psychol Human. 1988;14(2):281–94. https://doi.org/10.1037/0096-1523.14.2.281.
43. Wallsten TS, Budescu DV, Zwick R, Kemp SM. Preferences and reasons for communicating probabilistic information in verbal or numerical terms. B Psychonomic Soc. 1993;31(2):135–8. https://doi.org/10.3758/BF03334162.
44. Witteman CLM, Renooij S, Koele P. Medicine in words and numbers: a cross-sectional survey comparing probability assessment scales. BMC Med Inform Decis Making. 2007;7:13. https://doi.org/10.1186/1472-6947-7-13.
45. Reiss E. In quest of certainty. Am J Med. 1984;77(6):969–71. https://doi.org/10.1016/0002-9343(84)90170-0.
46. Renooij S, Witteman CLM. Talking probabilities: Communicating probabilistic information with words and numbers. Int J Approx Reason. 1999;22(3):169–94. https://doi.org/10.1016/S0888-613X(99)00027-4.
47. Hamm RM. Selection of verbal probabilities: A solution for some problems of verbal probability expression. Organ Behav Hum Dec. 1991;48(2):193–223. https://doi.org/10.1016/0749-5978(91)90012-I.
48. Budescu DV, Wallsten TS. Processing linguistic probabilities: General principles and empirical evidence. Psychol Learning Motivation. 1995;32:275–318. https://doi.org/10.1016/S0079-7421(08)60313-8.
49. Wallsten TS, Budescu DV. A review of human linguistic probability processing: General principles and empirical evidence. Knowl Eng Rev. 1995;10(1):43–62. https://doi.org/10.1017/S0269888900007256.
50. Witteman CLM, Renooij S. Evaluation of a verbal–numerical probability scale. Int J Approx Reason. 2003;33(2):117–31. https://doi.org/10.1016/S0888-613X(02)00151-2.
51. Robertson B, Vignaux GA, Berger CEH. Interpreting evidence: Evaluating forensic science in the courtroom, 2nd edn. Chichester, UK: John Wiley & Sons, 2016;92.
52. Hinkle DE, Wiersma W, Jurs SG. Applied statistics for the behavioral sciences. Boston, MA: Houghton Mifflin, 2003.
53. Chumbley LS, Morris MD, Kreiser MJ, Fisher C, Craft J, Genalo LJ, et al. Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm. J Forensic Sci. 2010;55(4):953–61. https://doi.org/10.1111/j.1556-4029.2010.01424.x.
54. Association of Forensic Science Providers. Standards for the formulation of evaluative forensic science expert opinion. Sci Justice. 2009;49(3):161–4. https://doi.org/10.1016/j.scijus.2009.11.004.
55. Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A model for case assessment and interpretation. Sci Justice. 1998;38(3):151–6. https://doi.org/10.1016/S1355-0306(98)72099-4.
56. Ott D, Thompson R, Song J. Applying 3D measurements and computer matching algorithms to two firearm examination proficiency tests. Forensic Sci Int. 2017;271:98–106. https://doi.org/10.1016/j.forsciint.2016.12.014.
57. Fadul TG, Hernandez GA, Stoiloff S, Gulati S. An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides. Miami, FL: Miami-Dade Police Department Crime Laboratory, 2011. Report No.: 237960.
58. Keisler MA. Isolated pairs research study. AFTE J. 2018;50(1):56–8.
59. Towler A, White D, Ballantyne K, Searston RA, Martire KA, Kemp RI. Are forensic scientists experts? J Appl Res Mem Cogn. 2018;7(2):199–208. https://doi.org/10.1016/j.jarmac.2018.03.010.
60. Saks MJ, Risinger DM, Rosenthal R, Thompson WC. Context effects in forensic science: A review and application of the science of science to crime laboratory practice in the United States. Sci Justice. 2003;43(2):77–90. https://doi.org/10.1016/s1355-0306(03)71747-x.
61. Stoel RD, Berger CEH, Kerkhoff W, Mattijssen EJAT, Dror IE. Minimizing contextual bias in forensic casework. In: Strom KJ, Hickman MJ, editors. Forensic science and the administration of justice: Critical issues and directions. Thousand Oaks, CA: SAGE Publications, Inc., 2014;67–86.
62. Dror IE. Human expert performance in forensic decision making: Seven different sources of bias. Aust J Forensic Sci. 2017;49(5):541–7. https://doi.org/10.1080/00450618.2017.1281348.
63. Dror IE. Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. Anal Chem. 2020;92(12):7998–8004. https://doi.org/10.1021/acs.analchem.0c00704.
64. Dror IE, Langenburg G. "Cannot decide": The fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide. J Forensic Sci. 2019;64(1):10–5. https://doi.org/10.1111/1556-4029.13854.
65. Montani I, Marquis R, Egli Anthonioz N, Champod C. Resolving differing expert opinions. Sci Justice. 2019;59(1):1–8. https://doi.org/10.1016/j.scijus.2018.10.003.
66. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. The elicitation of probabilities. Uncertain judgements: Eliciting experts' probabilities. Chichester, UK: John Wiley & Sons Ltd, 2006;61–96.
67. Lichtenstein S, Fischhoff B, Phillips LD. Calibration of probabilities: the state of the art to 1980. In: Kahneman D, Slovic P, Tversky A, editors. Judgment under uncertainty: Heuristics and biases. Cambridge, UK: Cambridge University Press, 1982;306–34.
68. Beach LR, Braun GP. Laboratory studies of subjective probability: a status report. In: Wright G, Ayton P, editors. Subjective probability. Chichester, UK: John Wiley and Sons, 1994;107–27.
69. Croskerry P, Norman G. Overconfidence in clinical decision making. Am J Med. 2008;121(5):S24–9. https://doi.org/10.1016/j.amjmed.2008.02.001.
70. Remus W, O'Conner M, Griggs K. Does feedback improve the accuracy of recurrent judgmental forecasts? Organ Behav Hum Dec. 1996;66(1):22–30. https://doi.org/10.1006/obhd.1996.0035.

71. Stone ER, Opel RB. Training to improve calibration and discrimination: the effects of performance and environmental feedback. Organ Behav Hum Dec. 2000;83(2):282–309. https://doi.org/10.1006/obhd.2000.2910.
72. Lichtenstein S, Fischhoff B. Training for calibration. Organ Behav Hum Perf. 1980;26(2):149–71. https://doi.org/10.1016/0030-5073(80)90052-5.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Mattijssen EJAT, Witteman CLM, Berger CEH, et al. Firearm examination: Examiner judgments and computer-based comparisons. *J Forensic Sci*. 2021;66:96–111. https://doi.org/10.1111/1556-4029.14557