Forensic Science International 313 (2020) 110331

Contents lists available at ScienceDirect

Forensic Science International

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

The interpretation of forensic conclusions by criminal justice professionals: The same evidence interpreted differently

Elmarije K. van Straalen[a,b,*], Christianne J. de Poot[a,b,c], Marijke Malsch[d], Henk Elffers[d]

[a] Amsterdam University of Applied Sciences, Forensic Sciences, P.O. Box 1025, 1000 BA Amsterdam, the Netherlands
[b] VU University Amsterdam, Criminology Department, De Boelelaan 1105, 1081 HV Amsterdam, the Netherlands
[c] Police Academy of the Netherlands, Apeldoorn, the Netherlands
[d] Netherlands Institute for the Study of Crime and Law Enforcement NSCR, Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Received 17 October 2019
Received in revised form 4 May 2020
Accepted 7 May 2020
Available online 13 May 2020

Keywords:
Forensic conclusions
Criminal justice professionals
Evidence interpretation
Strength of evidence
Communicating uncertainty

ABSTRACT

Forensic reports use various types of conclusions, such as a categorical (CAT) conclusion or a likelihood ratio (LR). In order to correctly assess the evidence, users of forensic reports need to understand the conclusion and its evidential strength. The aim of this paper is to study the interpretation of the evidential strength of forensic conclusions by criminal justice professionals. In an online questionnaire 269 professionals assessed 768 reports on fingerprint examination and answered questions that measured self-proclaimed and actual understanding of the reports and conclusions. The reports entailed CAT, verbal LR and numerical LR conclusions with low or high evidential strength and were assessed by crime scene investigators, police detectives, public prosecutors, criminal lawyers, and judges. The results show that about a quarter of all questions measuring actual understanding of the reports were answered incorrectly. The CAT conclusion was best understood for the weak conclusions, the three strong conclusions were all assessed similarly. The weak CAT conclusion correctly emphasizes the uncertainty of any conclusion type used. However, most participants underestimated the strength of this weak CAT conclusion compared to the other weak conclusion types. Looking at the self-proclaimed understanding of all professionals, they in general overestimated their actual understanding of all conclusion types.

1. Introduction

In the process of investigating and proving a crime, many different types of evidence may play a role, including witness statements, DNA, digital traces, fingerprints, observations, and shoeprints. It is impossible, and undesirable, for all professionals working with or making decisions based on this evidence to have expertise in all these different areas. In the forensic domain, there is a different expert for every type of evidence who will examine it and describe his or her findings and conclusion(s) in a report that is used by the professionals in the criminal justice system. For adequate functioning of the criminal justice process, three elements are important: (1) the report is clear, sound and correct, (2) the professional has a basic understanding of the content of the evidence and the report, and (3) the professional is able to correctly assess the evidential strength of the conclusion.

The first element actually means that every expert who investigates the same traces would come up with the same results. The study by De Keijser et al. [1] shows, however, that there can be a discrepancy in the content and layout of forensic reports of different experts. They asked 19 DNA experts around the world from 13 different forensic laboratories in seven countries to evaluate the same mixed DNA profiles and case information and write a DNA report in the way they usually do in their actual practice. The resulting reports differed in size, explanation of technical issues, use of explanation in appendices, evidence level of the proposition, use of contextual information, and the type and content of the conclusions. This implies that reports on the same DNA mark may be strikingly different and may yield different conclusions. It also gives a strong indication of the inherent uncertainties of incomplete and mixed DNA profiles [1]. In a study by Dror and Hampikian [2] 17 DNA experts working in the same forensic laboratory were presented with mixture DNA and the DNA profiles of suspects from a real adjudicated criminal case. Based on the presented evidence, the 17 independent DNA examiners varied in their conclusions from 'cannot be excluded' to 'excluded' and 'inconclusive'. Both studies [1,2] show that the same DNA evidence can be assessed differently by DNA experts, regardless of the

* Corresponding author at: Amsterdam University of Applied Sciences, Forensic Sciences, P.O. Box 1025, 1000 BA Amsterdam, the Netherlands
E-mail address: e.k.van.straalen@hva.nl (E.K. van Straalen).

http://dx.doi.org/10.1016/j.forsciint.2020.110331

forensic laboratory they are working in. For fingerprint evidence, research has shown possible bias in experts' conclusions as well [3,4]. An example of a mistaken identification by a fingerprint expert is the Mayfield case. On March 11, 2004, several bombs exploded on commuter trains in Madrid, eventually killing 191 people and wounding over 200. The fingermarks left on a bag with detonators connected with the attack were believed to match the fingerprints of Brandon Mayfield. This fingerprint identification was made by three members of the FBI Laboratory's Latent Print Unit. After Mayfield had been detained for two weeks, the Spanish National Police arrested an Algerian national whom they identified as the source of the fingermarks. He turned out to be 'a better match' than Mayfield [5]. The above studies show that findings of forensic investigations can be interpreted differently. Research therefore is necessary to gain better understanding of why experts can draw different conclusions about the same evidence. The users of forensic reports need to be aware of this human factor in decision-making. Besides the assumption that reports are correct, it is also important that a report clearly states what context information was used for the assessment and the conclusion, and what context information was left out of the assessment. By describing this, it is possible for the users of the reports to make a thought-through decision based on all available evidence.

For the second element, professionals have to be able to assess the meaning of the evidence in the forensic report. If they lack the necessary background to understand a forensic report, they may undertake training or gain advice from an independent source. Several studies [6–8] have shown that criminal justice professionals have a quite high level of self-proclaimed understanding of forensic reports, which is higher than their actual understanding. To understand that they need to seek training or advice, professionals need to be aware of their lack of knowledge. Besides understanding the content of a report, professionals also have to be aware of the possible subjectivity of experts assessing the evidence and writing the report and conclusion.

The third element is the assessment of the evidential strength of the conclusion. Do professionals have a correct understanding of forensic conclusions? Do they understand the evidential strength and the degree of uncertainty that are expressed in forensic conclusions? Do different types of professionals have the same understanding of these forensic conclusions? Finally, do different formulations that can be used to verbalise forensic evidence actually have the same meaning?

The main focus in this article is this third element, the understanding of forensic conclusions by criminal justice professionals. It is valuable to know whether the evidential strength of different formulations of the same conclusion is assessed in the same way by different professionals in the criminal justice system. Additionally, this paper examines professionals' self-proclaimed understanding of forensic conclusions.

1.1. Understanding the (un)certainty and evidential value of forensic conclusions

Forensic conclusions should always entail some degree of uncertainty, because no absolute certainty about the uniqueness of traces exists. To illustrate this, we will use fingerprint evidence as an example. Although we have some knowledge of the factors that influence the development of fingerprints [9], we do not know exactly how friction ridge patterns originate. Since it is practically impossible to compare all the fingerprints of the entire world's population, there can never be 100% certainty about the unicity of fingerprints. Furthermore, fingermarks found at a crime scene are hardly ever of perfect quality and quantity. A fingerprint examination is therefore usually based on the characteristics of only part of a fingermark, while the characteristics of the missing part remain unknown. The Organization of Scientific Area Committees for Forensic Science (OSAC) has drafted guidelines for fingerprint examination conclusions in which they state "A conclusion shall not be communicated as a fact. It is an interpretation of observations made by the examiner and shall be expressed as an expert opinion", [10]. Uncertainty should be expressed in all forensic conclusions. When it comes to the interpretation of the uncertainty expressed in forensic conclusions, several types of mistakes can be made. Koehler [11] provides a detailed description on possible fallacies within forensic decision making. We will discuss a few of these fallacies here. The prosecutor's fallacy and the defence fallacy are two common mistakes in the judgement of probabilities [6,12]. We will explain these fallacies using the likelihood ratio (LR), which represents the expert's view of the relative probability of the observed features of a trace under alternative hypotheses about the source of the trace.

An expert states:
'There were 12 corresponding minutiae between the fingermark and the reference fingerprint of suspect B. Smith, no differences were found. Two hypotheses have been formulated.
Hypothesis 1: The fingermark originated from the suspect.
Hypothesis 2: The fingermark originated from an unknown person.
The findings of this examination are 5 million times more probable when Hypothesis 1 is true versus when Hypothesis 2 is true'.

Prosecutor's fallacy → It is 5 million times more probable that the fingermark belongs to the suspect.
Correct understanding → The findings of the comparison are 5 million times more probable when the fingermark belongs to the suspect.

Defence fallacy → Since the total population is 17 million, the chances are greater that the fingermark is from someone else in the population than that it is from the suspect.
Correct understanding → Not all of these 17 million people have the same chances of leaving a fingermark at that specific crime scene (those who have never been near the crime scene or are physically not capable of entering the crime scene).

Another possible mistake in the interpretation of the evidence occurs when evidence is incorrectly used to answer evidential questions. When evaluating evidence, a certain hierarchy of propositions is followed to distinguish the source level, activity level, and offence level. Information about the source level is insufficient to answer questions about the offence level [13–15].[1] For example, the fingerprint examination conclusion about suspect B. Smith provides only information about the source level as to whether the fingermark might belong to the suspect. Although the conclusion provides information about the likelihood that an object is touched by the suspect, this conclusion does not provide any information about the activity or crime relatedness of the fingermark. More evidence is needed to draw conclusions about the activity level (did B. Smith perform a certain activity?) or offence level (did B. Smith commit the crime?). In an article by De Ronde et al. [16] the difference between source and activity level in relation to fingermark evidence is more profoundly explained.

1.2. Interpretation of forensic reports and conclusions

What is known about the interpretation of forensic conclusions in practice? Several studies have been conducted on this topic.

For conclusions expressed with a verbal descriptor, the choice of words can influence its comprehension. In a study of the interpretation of probability phrases, Willems et al. [17] asked participants to assess various phrases by assigning each a point

[1] In the current study, we only focus on the source level: on the identification of the source of a fingermark.

estimate on a 0–100% scale. The results show that the interpretation differs greatly among the participants. For more extreme words such as 'always' and 'never', this difference is the smallest [17]. Mullen et al. [18] asked students to point out the strength of a forensic statement containing one of ten verbal scale descriptors, such as 'weak support', 'moderate support', and 'extremely strong support'. The majority did not correctly understand the meaning of the terms used in the verbal scale. They attributed more value than they should to the weaker verbal conclusions and less value than they should to the stronger conclusions. Carter et al. [19] examined the utility of additional conclusions to the current categorical evidence scale used for fingerprint examinations as was proposed by the Friction Ridge Subcommittee of the OSAC. This committee suggests using a 5-conclusion scale, adding the conclusions 'support for different sources' and 'support for common sources' to the current 'exclusion', 'inconclusive', and 'identification'. The study shows that fingerprint examiners used these additional 'support for' conclusions about 35% of the time, which supports the need for more qualified conclusions for fingerprint comparisons. Carter et al. [19] discuss the theoretical difference between using categorical conclusions and likelihood ratio conclusions for fingerprint examinations. In a study by Arscott et al. [20] participants were asked to read a brief case summary that included a shoeprint evidence statement containing a verbal expression of the comparison conclusion and to rate the perceived strength of the expression. The results showed that when the numerical values were left out, both professionals and lay person perceived the three highest gradations in the verbal scale similarly. In general, the stronger the actual evidential strength, the stronger the participants assumed the evidential strength to be [20]. These studies show that the interpretation of the evidential value of verbal descriptions of probabilities can vary among those assessing them, and that adding verbal conclusions to existing scales can result in a rather different evidential value chosen by forensic experts.

In reaction to Arscot et al. [20], Berger and Stoel [21] wondered whether it is necessary to use only verbal expressions when there is the LR to accompany it. Wintle et al. [22] investigated whether the understanding of verbal probability expressions could be supported by providing numbers alongside these expressions. Lay participants read statements containing verbal expressions of probabilities. The numerical probabilities of the verbal expressions (very unlikely, unlikely, likely, very likely) were presented in three different ways: in a table in a separate browser window, in a tooltip that appeared when a mouse hovered over it, or between brackets behind the verbal expression. A fourth (control) group did not receive any numerical probabilities. After reading the statements, participants had to estimate the minimum, best, and maximum numerical probability belonging to the verbal expression. The results showed a high variability in the numerical probabilities assigned to the verbal probabilities. This variability was lowest when the numerical value was presented in brackets behind the verbal expression [22]. In a moot court exercise Langenburg et al. [23] explained and presented a likelihood ratio stating the strength of fingerprint evidence to a mock jury. The mock jurors were positive about the credibility of the evidence and understood most of the testimonies. However, fingerprint experts watching the moot court exercise were less positive about the use of a probability model for fingerprint examinations [23]. These studies show that verbal probability descriptions can be ambiguous and that numerical expressions can help reduce this ambiguity. When explained clearly, jurors can be positive about the use and understandability of numerical expressions. However, when using a 'new' type of conclusion in an existing field of expertise, it is important to not only focus on the users of the evidence, but also on the experts conducting the comparisons.

Are forensic conclusions assessed differently depending on the way they are expressed? In the study by De Keijser et al. [6], professionals assessed reports on facial comparison and DNA evidence in which the conclusion was presented in a verbally or visually expressed LR. The professionals did not assess the reports and conclusions significantly differently. Other studies did find differences in the interpretation of conclusions depending on the way they were expressed. McQuiston-Surrett and Saks [24] studied how jurors and judges assessed conclusions that were phrased differently. The study showed that jurors and judges considered the evidential value of conclusions to be higher if they contained the words 'match' or 'similar in all ... characteristics' or if they were described with an objective single-probability (the more qualitative conclusions) than if they were described with a subjective probability or an objective multi-frequency conclusion (the more quantitative conclusions) [24]. In a study by Martire et al. [25], lay participants received a brief case summary with the expert testimony of a fingerprint examiner about fingermarks found at the crime scene. The expert conclusion, which had a low or high evidential strength, was presented numerically, verbally in a table, or with a visual LR. The low evidential strength conclusion that was expressed with a verbal LR was deemed the weakest compared to all the other conclusions [25].

Thompson et al. [26] asked jury-eligible adults to assess pairs of comparisons entailing different phrasings of forensic conclusions on DNA or fingerprint examinations. They studied the interpretations of six different expression types with weak and strong evidential strength: LRs, strength of support statements, match frequencies and random match probabilities, likelihood of observed similarity, source probability statements, and categorical conclusions. In general, they found that statements designed to suggest that the strength of evidence was low or moderate were correctly perceived as weaker than statements designed to suggest that the strength of evidence was high. Conclusions phrased with the term 'match' were assessed as being extremely strong and assessed comparable to an 'LR of 10 million'. Conclusions phrased with the terms 'identification' were assessed as being considerably stronger than those phrased in terms of an 'individualisation', but these were considered significantly weaker than conclusions phrased in terms of 'match' or 'LR of 10 million'. Conclusions expressed with the phrasing 'extremely strong' were considered to be weaker than those using the term 'individualisation', even though this phrasing is intended to be comparable to an 'LR of 10 million' [26]. The different terms used to formulate the forensic conclusions were perceived differently than they were intended. Results of the study by Garrett, Michell, and Scurich (2018) show that jury eligible adults assessed categorical conclusions similar as match probabilities ranging from 1.000.000 to 10. Participants were similar in their decision to convict a suspect based on fingerprint evidence, regardless the conclusion type or specific strength of the evidence. Ratings of the likelihood the defendant left prints and committed the crime differed more between participants. Although those likelihoods were in general assessed as being slightly higher for the categorical conclusions, overall they were assessed similarly compared to the highest match probabilities. The highest presented probability (1.000.000) was assessed as having a significantly higher likelihood compared to the other probabilities, but the lower probabilities ranging from 100.000 to 10 were assessed similarly. Bayer et al. [27] studied the interpretation by jurors of different likelihood ratio presenting methods: only LR, LR with conversion table, LR with conversion table and figure relating prior and posterior probabilities, and match and nonmatch control groups. In general, participants overestimated the guilt of the suspect prior to the forensic evidence, and tended to undervalue the weight of the evidence. Most underestimation and overestimation of the guilt was found

when the match or nonmatch reporting technique was used. The use of likelihood ratios, provided with a conversion scale, a range of possible values, and instructions, seemed to help participants to evaluate the evidence.

Overall, it can be concluded that forensic verbal and numerical conclusions can be misinterpreted and that forensic conclusions reporting on the same evidence can be assessed differently depending on the way the evidence is expressed. This study will expand on this research. Since criminal justice professionals can be confronted with different conclusion types within forensic evidence, it is essential for them to know how to value these conclusions.

To investigate the understanding of different forensic conclusion types, we chose one forensic field of expertise in which different conclusion types are being used. In the Netherlands, most fingerprint examinations are conducted by the Dutch National Police. They express their conclusions of their fingerprint examinations in categorical terms. This conclusion type is used by most fingerprint examiners throughout the world. As illustrated above, in the past few years, new fingerprint comparison methods have been developed that use databases to calculate the frequency of certain fingermark characteristics. These frequencies are used to report on the evidential value of fingermarks in terms of numerical values such as likelihood ratios [28–30]. In the Netherlands, the Netherlands Forensic Institute (NFI) studies the occurrence rate for fingerprint patterns and details in a fingerprint database sample [31]. For the conclusions of their fingerprint examinations they use (verbal) likelihood ratios. Consequently, criminal justice professionals in the Netherlands can be confronted with multiple fingerprint examination reports using different conclusion types in one criminal case. This situation might be the future reality for other countries and for other forensic evidence fields.

Since professionals in the Netherlands are confronted with reports on fingerprint examinations using a categorical or likelihood ratio conclusion, we examine the interpretation of these different forensic conclusions in fingerprint examination reports. We compare categorical, verbal likelihood ratio, and numerical likelihood ratio conclusions with either low or high evidential strength. Based on former research, some overestimation of the understanding of conclusions in general is expected. We expect that participants have the highest self-proclaimed understanding of categorical conclusions compared to the conclusions using likelihood ratios. Based on the literature, we expect the numerical likelihood ratios to be better understood than the verbal (likelihood ratios and categorical) conclusions. We expect the categorical conclusions to be overestimated compared to the likelihood ratio conclusions.

2. Method

2.1. Design

In an online questionnaire, participants were asked to read three fingerprint examination reports. Fig. 1 shows the survey design. The participants were randomly allocated to one of two groups. Group 1 only received reports with weak evidential strength conclusions, and group 2 only received reports with strong evidential strength conclusions. The reports varied in conclusion type, and each participant received a report with a categorical conclusion, a report with a verbal LR conclusion, and a report with a numerical LR conclusion. These three reports for group 1 contained the weak conclusions: A: categorical—'cannot rule out', B: verbal LR—'moderate', and C: numerical LR—'LR of 50'. Group 2 contained the strong conclusions: D: categorical—'individualisation', E: verbal LR—'extremely strong', and F: numerical LR—'LR of 5 million'.

To control for an order-effect, the order in which the reports were presented was varied, resulting in six conditions: ACB, BAC, CBA, DFE, EDF, and FED. To equally distribute the professionals over the six conditions, every professional type had its own set of six conditions. The survey tool (www.qualtrics.com) randomly allocated the professionals over these conditions.

2.2. Participants

Participants were criminal justice professionals who in their work practice could be tasked with assessing forensic evidence conclusions. The participants voluntarily participated after a request was sent out via email within their organisation. In the email, a URL linked them to the survey for their group of professionals. The participants were randomly assigned to one of the six conditions.

A total of 269 criminal justice professionals participated in the online questionnaire: 59 crime scene investigators, 78 police detectives, 57 public prosecutors, 30 criminal lawyers, and 45 judges. Of those 269 participants, 43.5% was female and 56.5% male. The age of the total group was between 25 and 76 years old ($M = 45$, SD = 11). The professionals had up to 45 years of working experience in their profession ($M = 12$, SD = 9). Table 1 presents an overview of the background characteristics by professional type.

2.3. Experience with fingerprint evidence

In all, 74% ($N = 200$) of the participants had seen at least one fingerprint examination report of the police or a forensic institute before taking the survey. Of those participants, 62% ($N = 123$) judged the police reports were clear. All participants were asked about the ways in which they had gained knowledge about fingerprint examination: 44% ($N = 117$) reported taking a course on this topic, 10% ($N = 28$) had attended a conference or meeting on the topic, and 34% ($N = 92$) had read literature about the topic. Another 18% ($N = 48$) stated they had never gained knowledge of fingerprint examination.

2.4. Procedure

Via the URL in the invitation email, participants were directed to the questionnaire on the website of the survey tool (www.qualtrics.com).[2] A welcome message stated that for a study on the interpretation of forensic reports, they were invited to read three fingerprint examination reports and answer the appurtenant questions. On the next page, questions were asked about the participants' profession, age, gender, and education. After these questions, the first report was displayed. All three reports had exactly the same layout except for the conclusion part. The report was a simplified version of the police fingerprint examination report in The Netherlands. The only information on the one-page report was the registration numbers of the case and trace and basic personal information of an individual. No further information was provided about the fingermark, individual, crime (scene), or other evidence.

After each report, the exact same set of 17 questions appeared. These questions are similar to those used in other studies [6–8]. Six questions were about the alleged understanding of the report, and were five point Likert scale and open text questions. One Likert scale question and the two open text questions were eventually not used in this study. Eleven questions measuring actual understand-

[2] All study materials are in Dutch. Requests for an English translation of the vignettes and questionnaire can be sent to Elmarije van Straalen (e.k.van.straalen@hva.nl).

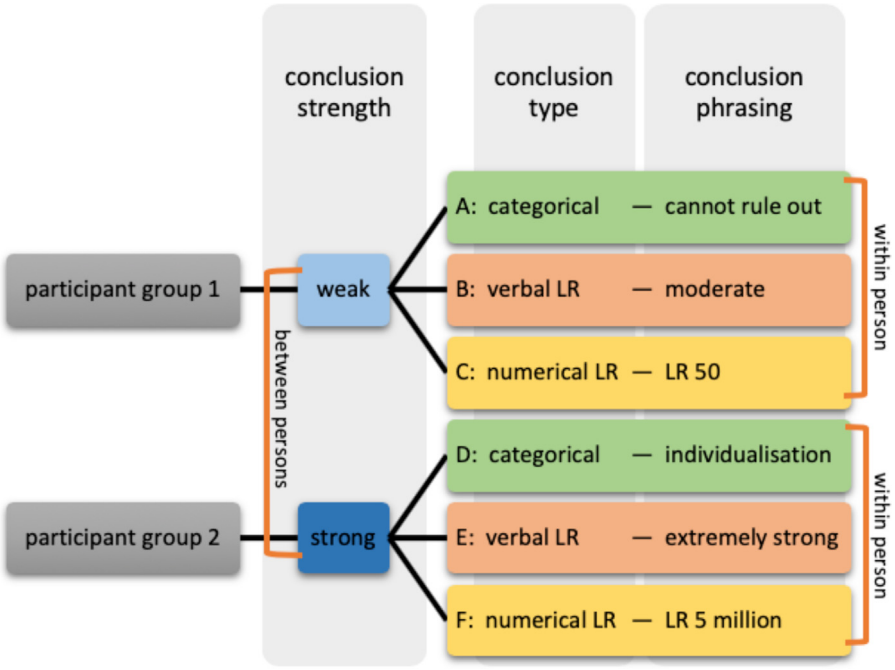

Fig. 1. Survey design.

Table 1
Background characteristics by professional type.

Professional type	Age (M)	Gender (% female/male)	Experience (M)
CSI	44 (SD = 12)	46/54	11 (SD = 8)
Police detective	47 (SD = 10)	36/64	14 (SD = 9)
Public prosecutor	44 (SD = 9)	51/49	12 (SD = 7)
Criminal lawyer	40 (SD = 12)	40/60	13 (SD = 9)
Judge	50 (SD = 11)	47/53	11 (SD = 11)

ing (see Tables 3 and 4) were about the assumed match between the fingermark and the fingerprint of the suspect, the assumed evidence against the suspect, and the assumed guilt of the suspect. These questions were five point Likert scale and single-answer multiple choice questions. Questions 1 and 3 (see Table 3) had the answer options 'yes', 'maybe', 'no', 'don't know'. Question 10 and 11 were five point Likert scale questions. The other questions had answer options 'yes', 'no', 'don't know'. The answers to questions 1–9 were recoded into 'correct' and 'incorrect'. In Table 3, for every question the answer that was recoded into 'correct' is between brackets.

Each report and each set of questions were presented on a new page. It was not possible to move to the next page without answering all the questions on that page, and returning to a previous page was prohibited. After the questions about the three reports were answered, a new set of questions was shown. These were questions about experience with fingerprint evidence and reports, and certain fingerprint procedures within the participants' organisation. At the end of the entire questionnaire, participants were asked for 'any comments on the questionnaire or reports you have read'. Most participants completed the study in 15–30 min.

In total, 378 professionals started the questionnaire, of which 109 stopped before they finished assessing the first report. Of those 378 participants, 269 finished assessing 768 reports. Twenty-six participants did not finish reading all three reports and answering all three sets of questions, only completing the first one or two. In the end, 241 participants finished answering all the questions, including the set of questions at the end of the questionnaire about experience and procedures. The data of the 269 participants were analysed.

2.5. Data analysis

All data were exported from the survey tool to IBM SPSS Statistics (version 23). The research design combines a within-person factor with a between-persons factor (Fig. 1). *Conclusion strength* is a between-persons factor with two levels: 'weak' and 'strong'. Participant group 1 only received weak conclusions, while participant group 2 only received strong conclusions. *Conclusion type* is a within-person factor with three levels: 'categorical', 'verbal LR', and 'numerical LR'. The phrasing of the conclusion depends on both conclusion type and strength. For weak conclusions, the phrasing of the categorical conclusion is 'cannot rule out' (A), for the verbal LR conclusion, it is 'moderate' (B), and for the numerical LR conclusion, it is 'LR of 50' (C). For strong conclusions, the phrasing of the categorical conclusion is 'individualisation' (D), for the verbal LR, it is 'extremely strong' (E), and for the numerical LR, it is 'LR of 5 million' (F). Comparing the effect of the stimuli within group 1 or within group 2 is a within-person comparison.

Because conclusion type and phrasing are nested within the strength factor, we compare A, B, and C in a separate set of tests from D, E, and F. We compare conclusion type and phrasing levels in pairs: A with B, A with C, and B with C, and in parallel, D with E, D with F, and E with F. For each comparison, we use a paired *t*-test. For comparing levels of the strength factor (between-persons), we use a two sample *t*-test. We first average results over the three nested type and phrasing results for the three reports that have been assessed by one person. In further analyses on how mean scores differ for different professionals, a *t*-test with Tukey post-hoc pairwise comparison has been used.

3. Results

The important question in this study is whether reports using different ways of presenting the conclusion while having a

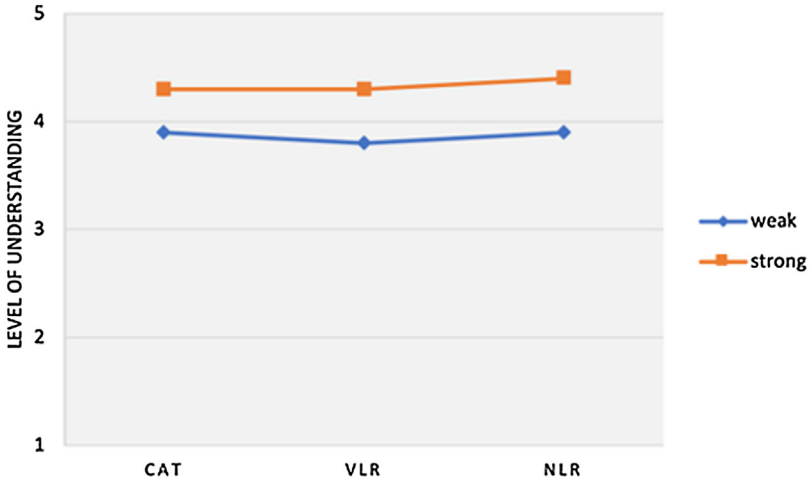

Fig. 2. The mean alleged understanding per type of conclusion for strong and weak evidential strength.
Note: Effect of conclusion strength significant at $p < .05$ level.

comparable evidential strength, are correspondingly assessed as being similar in evidential strength. Before looking at the actual understanding, we will look at the alleged comprehension of those conclusion types: do professionals think they understand the reports and conclusions?

3.1. Alleged comprehension

Participants were asked whether they understood the *conclusion* in the report. In general, they thought they understood the conclusion well. Fig. 2 shows the mean alleged understanding for the weak and strong conditions, divided into the three different conclusion types: CAT (categorical), VLR (verbal LR), and NLR (numerical LR). We observe no significant difference between the three conclusion types and phrasings, neither in the strong nor in the weak strength conditions, which is rather remarkable. The mean alleged understanding was significantly higher for conclusions with strong evidential strength than for the weak conclusions ($F(1,242) = 27.126$, $p < 0.001$). Therefore, we may conclude that it is the strength of the evidence, irrespective of the way it is expressed, that makes the participants confident about their understanding.

Participants were also asked whether they in general understood the *content* of the report. Overall, they thought they understood the content well ($M = 4.1$). For strong evidential strength reports, the alleged understanding of the content was significantly higher ($M = 4.3$) than for those with weak evidential strength ($M = 3.9$) ($F(1,242) = 21.410$, $p < 0.001$). The understanding of the *content* of the reports in general seems to be assessed similarly as the understanding of the *conclusion*. Indeed, both answers are highly correlated ($r = .847$, $p < 0.001$). In general, participants thought there was rather enough information in the report to understand the conclusion. Participants assessing reports with strong evidential strength seemed to be significantly more satisfied with the amount of information provided to understand the conclusion than participants assessing reports with weak evidential strength ($F(1,242) = 40.181$, $p < 0.001$). The alleged comprehension of the conclusion and report in general and the assumed amount of information provided to understand the conclusion did not differ significantly by conclusion type and phrasing.

Table 2 shows the mean alleged understanding per type of professional (with no distinction between evidential strength and conclusion type and phrasing). The answers to the questions on alleged comprehension differed significantly depending on the type of professional assessing the conclusion, $F(4,236) = 2.480$, $p < 0.05$, see Table 2. The CSIs had the highest alleged understanding of the conclusion, while the lawyers had the lowest. There was a significant difference in understanding of the content of the report between the various types of professionals (see Table 2), $F(4,236) = 2.879$, $p < 0.05$. Again, lawyers had the lowest alleged understanding of the reports, and CSIs had the highest alleged understanding. Also, the amount of information given in the report to understand the conclusion was assessed significantly differently by the different types of professionals, $F(4,236) = 2.611$, $p < 0.05$

Table 2
The mean alleged understanding per type of professional.

Questions ***	Type of professional					
	Total	CSI	Police detective	Public prosecutor	Lawyer	Judge
Do you understand the conclusion in the report? (1 I do not understand at all—5 I completely understand)	4.1*	4.3	4.0	4.2	3.8	4.0
Do you in general understand the content of the report? (1 I do not understand at all—5 I completely understand)	4.1*	4.4	4.0	4.2	3.9	4.0
Do you think there is enough information in the report to understand the conclusion? (1 not enough at all—5 completely enough)	3.4*	3.6[a]**	3.5[a]**	3.5[a]**	3.0[b]**	3.3[a,b]**

Note: *Effect of professional type significant at $p < 0.05$ level. **Means in the same row that do not share superscripts differ at $p < 0.05$. A Tukey post-hoc test revealed that the mean was significantly lower for the lawyers (2.96 ± 0.972) compared to the CSIs (3.58 ± 1.162, $p < 0.05$), police detectives (3.50 ± 1.145, $p < 0.05$), and public prosecutors (3.54 ± 1.065, $p < 0.05$). ***The answering scales are in parentheses.

Table 3
The percentage of correct answers per type of conclusion.

Questions and statements**	Total	Conclusion strength and type							
		Weak—group 1				Strong—group 2			
		Total weak	CAT (A)	VLR (B)	NLR (C)	Total strong	CAT (D)	VLR (E)	NLR (F)
1. Does the fingermark belong to the suspect? (maybe)	56[1]	75*	88[a]	64[b]	73[b]	37	33	41	36
2. Do you think it is impossible for the fingermark to be from someone else than the suspect? (no)	74[1]	83*	90[a]	78[b]	82[a,b]	64	56	70	67
3. It is ruled out that the fingermark belongs to someone else than the suspect. (no)	84[1]	88*	95[a]	83[b]	86[a,b]	79*	69[a]	84[b]	85[b]
4. The conclusion better fits the scenario that the fingermark belongs to the suspect than the scenario that it belongs to someone else. (yes)	80[1]*	65*	16[a]	86[b]	92[b]	95	92	96	96
5. There is more than a 50% chance the fingermark belongs to the suspect. (no)	26[1]*	43*	59[a]	31[b]	40[b]	9	8	9	10
6. The result of this examination is incriminating for the suspect. (yes)	72[1]*	63*	28[a]	73[b]	86[c]	81	80	83	81
7. The outcome of this examination is evidence against the suspect. (yes)	70[1]*	57*	27[a]	66[b]	78[b]	83	81	84	82
8. It has been proven that the defendant is guilty. (no)	89[1]	93	96	92	90	84	85	85	84
9. It has been proven that the suspect was at the scene were the fingermark was found. (no)	64[1]	73*	88[a]	66[b]	64[b]	57	53	59	59

Note: The means for weak or strong conclusion types in the same row that do not share the same superscripts differ at $p < 0.05$. *Effect of conclusion type significant at $p < 0.05$ level. **The correct answers to the questions are in parentheses. [1]Effect of conclusion strength significant at $p < 0.05$ level.

(see Table 2). For the alleged understanding of the conclusion and report in general, a Tukey post-hoc test showed no significant difference when the professional types were compared to each other. Overall, professionals thought they understood the report and conclusion well and thought there was rather enough information in the report to support this understanding.

3.2. Actual comprehension

Eleven questions measured the actual understanding of the conclusion. Table 3 shows the percentages of assessed reports with correct answers, and Table 4 shows the mean of the answers for questions with answering scales.

In general, most correct answers (89%) were given to the question about the guilt of the suspect (Q8). The fewest correct answers (26%) were given to the statement *'There is more than* 50% *chance the fingermark belongs to the suspect'* (Q5). This statement contains the prosecutor's fallacy, since in fact the reports and conclusions provide information about the *probability of finding the outcome* when the fingermark belongs to the suspect and not the *direct chance* the fingermark belongs to the suspect. For 68% of all the questions, the correct answers were given. When the outlier question about the 50% chance is deleted (Q5), it becomes 74%. In other words, the mean percentage of all questions answered incorrectly in all reports is 26%.

3.3. Effect of conclusion strength

For all questions, there was a statistically significant difference between groups with strong and groups with weak evidential strength conclusions as determined by a one-way ANOVA (for all questions, $p < 0.05$). Questions that were incorrectly answered by participants evaluating reports containing strong conclusions were mainly answered with an overestimation of the evidential strength. For example, to the question '*Does the fingermark belong to the suspect?'* (Q1), in 62% of the reports with a strong conclusion, the participant answered 'yes'. For the weak conclusions, this was 21% (the correct answer is 'possibly'). For six out of nine questions, there were more correct answers provided assessing reports with weak evidential strength conclusions compared to the ones with strong evidential strength. All six questions were about the ability to be absolutely certain about the match between a fingermark and a fingerprint.

3.4. Effect of conclusion type

Are different conclusion types and conclusion phrasings assessed differently? When we look at the effect of the type of conclusion on the understanding of the reports, a *t*-test shows this effect to be statistically significant for the questions '*How likely is it that the fingermark belongs to the suspect?'* (Q10) ($F(2,765) = 15,895$,

Table 4
The mean (M) answers per type of conclusion.

Questions and statements**	Total	Conclusion strength and type							
		Weak—group 1				Strong—group 2			
		Total weak	CAT (A)	VLR (B)	NLR (C)	Total strong	CAT (D)	VLR (E)	NLR (F)
10. How likely is it that the fingermark belongs to the suspect? (1 very unlikely—5 very likely)	4.03*	3.5*	2.9[a]	3.7[b]	3.9[b]	4.6	4.6	4.6	4.5
11. To what extent do you think the conclusion of the expert about the fingerprint evidence is incriminating or exculpatory for the suspect? (1 very exculpatory—5 very incriminating)	3.96*	3.5*	2.9[a]	3.7[b]	3.9[b]	4.4	4.4	4.5	4.4

Note: Means in the same row that do not share superscripts differ at $p < 0.05$. *Effect of conclusion type significant at $p < 0.001$ level. **The answering scales are in parentheses.

Table 5
Comparison of best alleged and best actual understood strength, conclusion type, and professional.

	Alleged	Actual
Best understood evidential strength	Strong	Weak
Best understood weak conclusion	Numerical LR	CAT
Best understood strong conclusion	Numerical LR	Verbal LR
Best understanding professional	CSI	Lawyer

$p < 0.001$), *'To what extent do you think the conclusion of the expert about the fingerprint evidence is incriminating or exculpatory for the suspect?'* (Q11) ($F(2,765) = 18,756$, $p < 0.001$), *'There is more than a 50% chance the fingermark belongs to the suspect'* (Q5) ($F(2,765) = 5,978$, $p < 0.01$), *'The outcome of this examination is evidence against the suspect.'* (Q7) ($F(2,765) = 23,883$, $p < 0.001$), *'The result of this examination is incriminating for the suspect'* (Q6) ($F(2,765) = 33,253$, $p < 0.001$), and *'The conclusion better fits the scenario that the fingermark belongs to the suspect than the scenario that it belongs to someone else'* (Q4) ($F(2,765) = 96,436$, $p < 0.001$). These effects seem to be caused by the weak CAT conclusion (A), which is phrased 'cannot rule out'. As Tables 3 and 4 show, the answers for reports containing this conclusion differ significantly from reports containing the other conclusions. The evidential strength of reports expressed in the weak CAT conclusion ('cannot rule out') is often underestimated when compared to the other conclusions. For example, to the statement *'The result of this examination is incriminating for the suspect'* (Q6) in only 28% of the reports containing the weak CAT conclusion (A), the participant gave the correct answer 'yes'. For reports with other conclusion types, more than 70% answered 'yes'. It can therefore be stated that the phrasing 'cannot rule out' (A), used to express the weak CAT conclusion, is assessed as being less incriminating for the suspect compared to other conclusions. For reports with a strong evidential strength conclusion (group 2), there is a significant effect of the conclusion type for the statement *'It is ruled out that the fingermark belongs to someone else than the suspect'* (Q3). Participants assessing reports with the strong CAT conclusion (D), phrased as 'individualisation', answered in 30% of the reports 'yes' (compared to 13% and 11% for the strong verbal LR (E) and strong numerical LR (F) conclusions) instead of the correct 'no'. Reports phrased with the strong CAT conclusion (D) thus seem to impart a stronger belief that the fingermark cannot belong to someone else than the person it is 'individualised' to than reports formulated with an LR do.

For reports with weak evidential strength conclusions (group 1), there is a significant effect of the type of conclusion for all questions (except for Q8, *'It has been proven that the defendant is guilty'*). This effect seems to be caused by the weak CAT conclusion, phrased as 'cannot rule out' (A), of which the evidential strength seems to be underestimated compared to the other conclusions with the same evidential strength. For example, to the statement *'The conclusion better fits the scenario that the fingermark belongs to the suspect than the scenario that it belongs to someone else'* (Q4), participants answered 'no' in 63% of the reports containing the weak CAT conclusion (A) (compared to 6% and 3% for the weak verbal and numerical LR conclusions (B and C)). The correct answer here is 'yes', which was only given in 16% of the reports with this conclusion type (compared to 86% and 92% for the weak verbal and numerical LR conclusions). Therefore, we can state that in general, participants felt that the weak CAT conclusion phrased as 'the suspect cannot be ruled out as donor of the fingermark' (A) does not best fit the scenario that the fingermark belongs to the suspect instead of the scenario that it does not belong to the suspect.

3.5. Alleged and actual comprehension

In general, participants thought they understood the conclusions and reports well (see Fig. 2). When it comes to the actual understanding for all groups, mistakes are made in the assessment of the evidential value of all conclusion types (see Table 3). Therefore, it can be stated that in general, there is an overestimation of comprehension. Table 5 shows an overview of the best alleged understood and best actual understood evidential strength, conclusion type, and phrasing and best alleged and actual understanding professional.

Looking at the evidential strength, participants who evaluated reports with strong evidential strength conclusions (group 2) had a higher alleged understanding of the report and conclusion compared to participants evaluating reports with weak evidential strength conclusions. When it comes to the actual understanding, for only three out of nine questions, the percentage of correct answers was highest for the reports with strong conclusions. It therefore seems that the actual understanding is highest for the reports with weak evidential strength conclusions and the

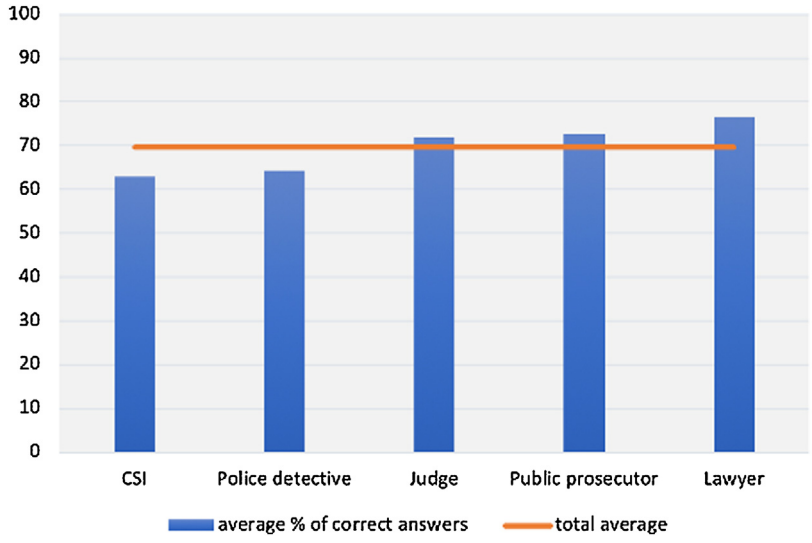

Fig. 3. The mean percentage of correct answers for all questions per type of professional.

Table 6
The mean (M) answers per type of professional.

Questions and statements (M)*	Total	Type of professional (M)				
		CSI	Police detective	Public prosecutor	Lawyer	Judge
10. How likely is it that the fingermark belongs to the suspect? (1 very unlikely—5 very likely)	4.03	4.16	4.17	3.91	3.97	4.04
11. To what extent do you think the conclusion of the expert about the fingerprint evidence is incriminating or exculpatory for the suspect? (1 very exculpatory—5 very incriminating)	3.96	4.11	3.95	3.91	3.89	3.93

Note: *The answering scales are in parentheses.

overestimation of understanding is highest for the strong evidential strength conclusions and reports.

For the alleged comprehension of the different conclusion types with weak evidential strength (group 1), participants seemed to think they best understood the numerical LR report (C) and least understood the report with a verbal LR conclusion (B). For the actual understanding, for only three out of nine questions, most correct answers were given for reports with the numerical LR conclusion (C), and none of the questions had most correct answers for the verbal LR conclusion (B) (when comparing the weak conclusions). Looking at the actual understanding of reports with weak CAT conclusions (A), for six out of nine questions, participants understood these reports the best, making this the best understood weak conclusion type.

The alleged comprehension of strong evidential strength conclusions (group 2) was highest for the numerical NLR conclusion (F), while for reports with this conclusion type, only two out of nine questions had the most correct answers. The CAT conclusion for strong evidence (D) was expected to be least understood, and since for eight out of nine questions the percentage of correct answers was lowest for this conclusion type, this seems to be the conclusion type with least overestimation of comprehension. The verbal LR conclusion (E) had the highest actual understanding considering the most correct answers for seven out of nine questions, while the alleged understanding of this conclusion type was average compared to the other conclusion types. As stated, the alleged understanding of conclusion types does not differ significantly. However, the actual understanding does, which shows the professionals' lack of awareness of their actual understanding of different forensic conclusions.

Of all the participating professionals, the CSIs were most confident about their understanding of the reports and its conclusions. Looking at the actual understanding, the CSIs had for the majority of the questions the lowest or second lowest percentage of correct answers. Criminal lawyers were least optimistic about their understanding but had the highest or second highest percentage of correct answers to all questions except for the question about whether the results are incriminating for the suspect, which might be explained by the lawyers' professional partiality towards the defendant.

3.6. Participant characteristics

Fig. 3 shows the mean percentages of correct answers to all questions per type of professional for all assessed reports. Table 6 shows the mean of the answers with answering scales per professional type. For all the questions, the type of professional answering the questions had a significant effect on the percentage of correct answers.

For the total group, the gender of the participants had no significant effect on the percentage of correct answers to all questions. However, a one-way ANOVA showed that for the total group of participants, their age had a significant effect on the answer to the statement '*It is ruled out that the fingermark belongs to someone else than the suspect*' (Q3) ($F(4,238) = 2,793$, $p < 0.05$). In general, most correct answers were given by participants between 30 and 39 years old ($N = 61$, $M = 1,09$, SD = 0,20) and most incorrect answers by participants older than 60 ($N = 26$, $M = 1,31$, SD = 0,39).

Whether participants had seen a fingerprint examination report or gained knowledge about fingerprint examination before completing the questionnaire had no effect on the actual comprehension of the conclusions. Participants were asked how familiar they feel when it comes to reading and understanding figures and statistics. Overall, they felt quite familiar ($M = 3.5$). In general, non-legal professionals (CSI's and police detectives) felt slightly more familiar with statistics ($M = 3.6$) than legal professionals (public prosecutors, criminal lawyers, and judges) ($M = 3.3$). As shown by Fig. 3, legal professionals had a higher percentage of correct answers compared to non-legal professionals.

For all professional groups, the differences per type of conclusion were similar for all questions. We examined for every group of professionals whether they answered questions significantly differently per type of conclusion. For most questions, we saw similar trends. For the weak evidential strength, the CAT (A) conclusion was significantly different from the weak numerical LR (C) and mostly from the weak verbal LR (B) conclusion as well.

4. Discussion and conclusion

This study focused on the interpretation of forensic conclusions by criminal justice professionals. Do professionals understand the forensic conclusions they work with in daily practice and on which important decisions for the criminal justice system are based? Overall, professionals *thought* they understood the reports and their conclusions quite well. Those with the lowest alleged understanding had the highest actual understanding, whereas participants with the highest alleged understanding had (almost) lowest actual understanding. Some overestimation of the self-proclaimed understanding was expected, but not this skewed association between alleged and actual understanding. These outcomes show that professionals do not always correctly assess their understanding of forensic conclusions. This is in line with findings by other studies [6–8].

Looking at the actual understanding, participants seemed to overestimate the strength of almost all conclusion types and phrasings, except for the weak CAT conclusion, which was undervalued by most participants. As expected [20,26,32], reports with strong conclusions were assessed as having a higher evidential strength than reports with weak conclusions. The results showed that the strength of reports with strong conclusions was most often overvalued. Participants evaluate those reports as having a higher evidential strength than they actually do, allocating them 100% certainty. Based on the literature [26], we expected the numerical LR conclusion to be better understood than the verbal (verbal LR and CAT) conclusions, but in fact, it was the

other way around: for the weak conclusions, the CAT conclusion was best understood, and for the strong conclusions, the verbal LR conclusion was best understood. When the different conclusion types are compared, the CAT conclusion stands out. This conclusion type seems to be misinterpreted the most. Compared to the other conclusion types, the strong CAT conclusion is often overvalued (as was also seen in the study by Garrett et al. [32]), whereas the weak CAT conclusion is mostly undervalued. These findings of over- and underestimation of the strength of the CAT conclusions are similar to results in the study by Bayer et al. [27]. The weak CAT conclusion in particular is valued differently, being assessed as less incriminating for the suspect compared to the weak verbal LR and the weak numerical LR conclusions.

The various types of professionals assessed the reports and conclusions differently. Can this difference be explained by the specific roles they have in the criminal justice system? The CSI collects the evidence and passes it on to the police detective and public prosecutor. The CSI does not have to make a decision based on the conclusion, but often writes a summary of the report and its conclusion. The police detective has to decide whether and how to use the evidence for the investigation. The public prosecutor decides based on the conclusion whether to use it in the criminal case and trial. The criminal lawyer has to decide based on the report and conclusion whether the evidence was used correctly and in accordance with the law and whether there is room for a defence strategy. And finally, the judge has to decide based on the report and conclusion whether it is sufficient evidence and in what direction it points (guilt or innocence). Every professional has his or her own role in the judgment of the evidence. The role of the CSI (making a summary and passing it on) is different than the role of the judge (making a final and 'absolute' decision). We cannot deny the difference in the way professionals use forensic conclusions and the possibility that this influences their interpretation. Also, differences in their familiarity with statistics might influence their interpretation of those conclusions. Further research is needed to explain individual differences and effects of participants' background characteristics. Still, we think every professional working with forensic conclusions should understand the meaning and value of these conclusions. Criminal justice professionals working with forensic evidence can make mistakes in the investigation and prosecution process when their understanding of forensic conclusions is not entirely correct. To enhance this understanding, what would be the best way to present a forensic conclusion?

It is often claimed that the strong CAT conclusion 'individualisation' is best suited for the traditional forensic examination reports, but this conclusion type and its phrasing seem to be the most misinterpreted and overvalued conclusion. We do not suggest abandoning this way of reporting results, but when verbal expressions of conclusions are used, the choice of words is determinative for its understanding. Conclusions entailing some sort of uncertainty seem to be better understood. Adding conclusions to an existing scale might be useful, especially if they provide more defined options for expressing examination results [19].

Based on our findings, we suggest that for every forensic conclusion, its uncertainty should be stated clearly. The words 'cannot be ruled out as donor' seem to make professionals aware of uncertainty in conclusions. Adding these words as a kind of 'disclaimer' to other conclusions might help professionals with their interpretation of the value of conclusions and the awareness that there is always some uncertainty. When using a numerical value, it has to be clear what the foundation for the numerical value is. If the investigation method is not scientifically based and a calculation for the choice of numerical value cannot be clearly stated, this might not be the best conclusion type. It is debatable whether this is the same when using a verbal scale, since any kind of scale might impart the assumption of a clear distinction between possible outcomes.

The communication on the evidential value and uncertainty of a specific piece of evidence should be clear enough for everyone using the same evidence to have the same understanding. In our study we observe differences in interpretation and therefore differences in understanding. To enhance the uniformity in this understanding we have made a few suggestions.

Based on former studies and on our findings, we suggest the use of a numerical conclusion (which can be a likelihood ratio or another numerical value) with a clear description of the foundation of that value and its uncertainty. This should only be used when a clear foundation (calculation) for the numerical value can be presented. When a clear calculation cannot be presented, a verbal value would be better suited. Also for a verbal conclusion, a clear foundation for the chosen outcome and the existing uncertainty should be presented. A verbal scale should only be used, when a clear distinction between possible outcomes can be provided. Even though our study solely focused on the interpretation of evidential strength and conclusion types, we believe that since misinterpretations are easily made, reports should clearly state on what information the conclusion is based. This way, professionals using the reports are able to make a correct assessment of the evidence and its evidential strength. Professionals using reports with numerical conclusions need to have sufficient mathematic knowledge (or gain advice) to fully understand these values and be able to assess uncertainty. When decisions are made on the use of new conclusion types within a field of forensic evidence, the experts conducting the forensic comparisons should be involved in this process. The study by Langenburg et al. [23] indicates that the perception of a new methodology and corresponding conclusion can be different for laypersons compared to forensic experts.

We believe that professionals working with forensic reports and conclusions, should be aware of a possible lack in their understanding, and ask for advice from an independent source. By sharing the results of the current as well as comparable experimental studies with these professionals, awareness of lack of understanding can be increased. Furthermore, implementing standard feedback on professionals' assessment of evidential value could help increase this awareness. In the Netherlands, forensic advisors with a university degree in forensic science are employed at the courts. These advisors can provide explanation on forensic evidence to judges when asked for. We believe this advice should be standard for all criminal cases entailing forensic evidence. Such an initiative should also be implemented into other organizations within the criminal justice system handling forensic evidence. As long as these advisers withhold from opinionizing on the evidence, this could be an important addition.

Further research on the interpretation of (new) forensic conclusions is necessary to develop new conclusions or instructions. Since this study shows that professionals are not fully aware of a possible lack in their understanding of forensic evidence, they might not seek for extra assistance (advice, courses) when needed. Therefore, courses on the interpretation of forensic evidence should be obligatory for all professionals working with these conclusions. Our recommendation is to make these courses part of the basic training for police, public prosecutors, lawyers, and judges. We think that these obligatory courses should at least include insights into the creation of forensic reports and conclusions by the forensic experts, a discussion of scientific research on the understanding of forensic reports, conclusions and common fallacies, and teaching on the foundation and formation of numerical values, (Bayesian) statistics and the uncertainty in forensic evidence.

5. Limitations

The experiment conducted in this study is not a representation of an actual criminal case. In our study, criminal justice professionals only received a report with information on the source level of one fingermark. In daily practice, professionals have information on all available evidence in a case. This additional information might influence their interpretation of forensic conclusions. Furthermore, in daily practice, a forensic report usually contains a table with the verbal descriptions (verbal LR) and matching numerical LRs. We decided to present the verbal and numerical LR separately to assess for each individually how they are interpreted.

We do not know if the outcomes of this study would have been different if the forensic reports and conclusions were presented in a real case. However, by conducting a controlled experiment, we can study a specific aspect of the criminal justice decision-making process. We did not study the interpretation of forensic (fingerprint) reports. This study tells us how professionals interpret forensic conclusions and assess their value. This is important for the construction of new conclusions in the development of instructions and for educational purposes.

Funding

This work was supported by the Taskforce for Applied Research of the Netherlands Organisation for Scientific Research (NWO), research grant no. 2014-01-124PRO.

CRediT authorship contribution statement

Elmarije K. van Straalen: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Christianne J. de Poot:** Conceptualization, Methodology, Formal analysis, Writing - review & editing, Supervision, Funding acquisition. **Marijke Malsch:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Henk Elffers:** Formal analysis, Writing - review & editing.

Declarations of competing interest

None.

References

[1] J.W. de Keijser, M. Malsch, E.T. Luining, M. Weulen Kranenbarg, D.J. Lenssen, Differential reporting of mixed DNA profiles and its impact on jurists' evaluation of evidence. An international analysis, Forensic Sci. Int. Genet. 23 (2016) 71–82, doi:http://dx.doi.org/10.1016/j.fsigen.2016.03.006.
[2] I.E. Dror, G. Hampikian, Subjectivity and bias in forensic DNA mixture interpretation, Sci. Justice 51 (4) (2011) 204–208.
[3] I.E. Dror, C. Champod, G. Langenburg, D. Charlton, H. Hunt, R. Rosenthal, Cognitive issues in fingerprint analysis: inter- and intra-expert consistency and the effect of a 'target' comparison, Forensic Sci. Int. 208 (1-3) (2011) 10–17, doi:http://dx.doi.org/10.1016/j.forsciint.2010.10.013.
[4] I.E. Dror, S.M. Kassin, J. Kukucka, New application of psychology to law: improving forensic evidence and expert witness contributions, J. Appl. Res. Mem. Cogn. 2 (2013) 78–81, doi:http://dx.doi.org/10.1016/j.jarmac.2013.02.003.
[5] Office of the Inspector General, U.S. Department of Justice, A Review of the FBI's Handling of the Brandon Mayfield Case, U.S. Department of Justice, Washington, D.C, 2006.
[6] J.W. de Keijser, H. Elffers, R.M. Kok, M.J. Sjerps, Bijkans Begrepen: Feitelijk En Vermeend Begrip Van Forensische Deskundigenrapportages Onder Rechters, Advocaten En Deskundigen, Boom Juridische uitgevers, Den Haag, 2009.
[7] J.W. de Keijser, H. Elffers, Understanding of forensic expert reports by judges, defense lawyers and forensic professionals, Psychol. Crime Law 18 (2) (2012) 191–207, doi:http://dx.doi.org/10.1080/10683161003736744.
[8] M. Malsch, M.D. Taverne, H. Elffers, J.W. de Keijser, P.R. Kranendonk, DNA-rapporten: Makkelijker Kunnen We Het Niet Maken, Begrijpelijker Wel, Boom Lemma Uitgevers, Den Haag, 2013.
[9] C. Champod, C.J. Lennard, P. Margot, M. Stoilovic, Fingerprints and Other Ridge Skin Impressions, 2nd ed., CRC Press, Boca Raton, 2016.
[10] OSAC, Friction Ridge Skin Subcommittee, Standard for Friction Ridge Examination Conclusions (DRAFT), (2019) .
[11] J. Koehler, Error and exaggeration in the presentation of DNA evidence at trial, Jurimetrics J. 34 (1) (1993) 21–40.
[12] W.C. Thompson, E.L. Schumann, Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy, Law Hum. Behav. 11 (3) (1987) 167–187.
[13] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, A hierarchy of propositions: deciding which level to address in casework, Sci. Justice 38 (4) (1998) 231–239, doi:http://dx.doi.org/10.1016/s1355-0306(98)72117-3.
[14] G. Jackson, S. Jones, G. Booth, C. Champod, I.W. Evett, The nature of forensic science opinion—a possible framework to guide thinking and practicce in investigation and in court proceedings, Sci. Justice 46 (1) (2006) 33–44, doi:http://dx.doi.org/10.1016/s1355-0306(06)71565-9.
[15] B. Kokshoorn, B. Aarts, Tde Blaeij, P. Maaskant-van Wijk, B. Blankers, Bewijskracht van onderzoek naar biologische sporen en DNA: deel 1. Theoretisch kader en aandachtspunten bij conclusies in het deskundigenrapport, Expert. En Recht 6 (2014).
[16] A. de Ronde, B. Kokshoorn, C.J. de Poot, M. de Puit, The evaluation of fingermarks given activity level propositions, Forensic Sci. Int. 302 (2019) 109904, doi:http://dx.doi.org/10.1016/j.forsciint.2019.109904.
[17] S.J.W. Willems, C.J. Albers, I. Smeets, Variability in the Interpretation of Dutch Probability Phrases - a Risk for Miscommunication arXiv:1901.09686 [stat.OT], (2019) .
[18] C. Mullen, D. Spence, L. Moxey, A. Jamieson, Perception problems of the verbal scale, Sci. Justice 54 (2) (2014) 154–158, doi:http://dx.doi.org/10.1016/j.scijus.2013.10.004.
[19] K.E. Carter, M.D. Vogelsang, J. Vanderkolk, T. Busey, The utility of expanded conclusion scales during latent print examinations, J. Forensic Sci. (2020), doi:http://dx.doi.org/10.1111/1556-4029.14298.
[20] E. Arscott, R. Morgan, G. Meakin, J. French, Understanding forensic expert evaluative evidence: a study of the perception of verbal expressions of the strength of evidence, Sci. Justice 57 (3) (2017) 221–227.
[21] C.E.H. Berger, R.D. Stoel, Response to "A study of the perception of verbal expressions of the strength of evidence", Sci. Justice 58 (1) (2018) 76–77.
[22] B.C. Wintle, H. Fraser, B.C. Wills, A.E. Nicholson, F. Fidler, Verbal probabilities: very likely to be somewhat more confusing than numbers, PLoS One 14 (4) (2019)e0213522, doi:http://dx.doi.org/10.1371/journal.pone.0213522.
[23] G. Langenburg, C. Neumann, S.B. Meagher, C. Funk, P.A. Julieanne, Presenting probabilities in the courtroom: a moot court exercise, J. Forensic Ident. 63 (4) (2013) 424–488.
[24] D. McQuiston-Surrett, M.J. Saks, The testimony of forensic identification science: what expert witnesses say and what factfinders hear, Law Hum. Behav. 33 (5) (2009) 436–453, doi:http://dx.doi.org/10.1007/s10979-008-9169-1.
[25] K.A. Martire, R.I. Kemp, M. Sayle, B.R. Newell, On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect, Forensic Sci. Int. 240 (2014) 61–68, doi:http://dx.doi.org/10.1016/j.forsciint.2014.04.005.
[26] W.C. Thompson, R. Hofstein Grady, E. Lai, H.S. Stern, Perceived strength of forensic scientists' reporting statements about source conclusions, Law Probab. Risk 17 (2) (2018) 133–155.
[27] D. Bayer, C. Neumann, A. Ranadive, Communication of statisically based conclusions to jurors-A pilot study, J. Forensic Ident. 66 (5) (2016) 405–427.
[28] H.J. Swofford, A.J. Koertner, F. Zemp, M. Ausdemore, A. Liu, M.J. Salyards, A method for the statistical interpretation of friction ridge skin impression evidence: method development and validation, Forensic Sci. Int. 287 (2018) 113–126, doi:http://dx.doi.org/10.1016/j.forsciint.2018.03.043.
[29] A.J. Leegwater, D. Meuwly, M. Sjerps, P. Vergeer, I. Alberink, Performance study of a score-based likelihood ratio system for forensic fingermark comparison, J. Forensic Sci. 62 (2) (2017) 626–640.
[30] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae, J. Forensic Sci. 52 (1) (2007) 54–64, doi:http://dx.doi.org/10.1111/j.1556-4029.2006.00327.x.
[31] A. de Jongh, A.R. Lubach, S.L. Lie Kwie, I. Alberink, Measuring the rarity of fingerprint patterns in the dutch population using an extended classification set, J. Forensic Sci. (2018), doi:http://dx.doi.org/10.1111/1556-4029.13838.
[32] B. Garrett, G. Mitchell, N. Scurich, Comparing categorical and probabilistic fingerprint evidence, J. Forensic Sci. 63 (6) (2018) 1712–1717, doi:http://dx.doi.org/10.1111/1556-4029.13797.