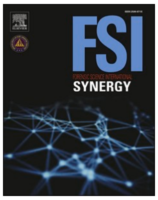
Forensic Science International: Synergy 7 (2023) 100437

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy



The interpretation of forensic conclusions by professionals and students: Does experience matter?

Elmarije K. van Straalen [a,b,*], Christianne J. de Poot [a,b,c], Marijke Malsch [d,e], 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
[e] Open University, Heerlen, the Netherlands

ARTICLE INFO

Keywords:
Forensic conclusions
Criminal justice professionals
Influence of experience
Evidence interpretation
Communicating uncertainty

ABSTRACT

Are professionals better at assessing the evidential strength of different types of forensic conclusions compared to students? In an online questionnaire 96 crime investigation and law students, and 269 crime investigation and legal professionals assessed three fingerprint examination reports. All reports were similar, except for the conclusion part which was stated in a categorical (CAT), verbal likelihood ratio (VLR) or numerical likelihood ratio (NLR) conclusion with high or low evidential strength. The results showed no significant difference between the groups of students and professionals in their assessment of the conclusions. They all overestimated the strength of the strong CAT conclusion compared to the other conclusion types and underestimated the strength of the weak CAT conclusion. Their background (legal vs. crime investigation) did have a significant effect on their understanding. Whereas the legal professionals performed better compared to the crime investigators, the legal students performed worse compared to crime investigation students.

1. Introduction

In the criminal justice system, crime investigators and legal professionals are tasked with assessing the content of forensic reports and interpreting their conclusions. It makes sense to assume that these professionals know how to interpret evidential values correctly due to their experience in evaluating the content of the forensic report. They have read and assessed forensic reports numerous times, and have expanded their experience with every report they dealt with. Nonetheless, we question whether this experience is always helpful. In general, experience provides a learning process that improves the skills and task performance of professionals, but does experience also help in making decisions about the reported value of forensic evidence? Previous studies have shown that criminal justice professionals can have difficulty interpretating forensic conclusions [1–4]. This raises the question of how students without professional experience interpret forensic conclusions, and whether this skill can be learned through experience. In this paper, we will test the influence of professional experience on the interpretation of forensic conclusions. We compare the performance of students learning to become crime investigators or legal professionals, with the performance of individuals already working in the forensic field as crime investigators or as legal professionals. Crime investigators participating in our study are police detectives and crime scene investigators; legal professionals are public prosecutors, criminal lawyers, and judges. We study how students and professionals interpret different types of forensic conclusions with comparable strength, to see whether professional experience enhances their performance. In this section, we discuss what we know from the literature on the interpretation of forensic conclusions (Section 1.1), followed by (Section 1.2) a literature description on how experience relates to the development of professional skills and to expertise[1] in general and how this learning process is influenced by theoretical knowledge and by feedback. Finally, we discuss the aim and hypotheses of the current study (Section 1.3).

1.1. The interpretation of forensic conclusions

Reports on forensic evidence usually describe specific case details and the results of a comparison made between a trace and reference

* Corresponding author. Amsterdam University of Applied Sciences, Forensic Sciences, P.O. Box 1025, 1000 BA Amsterdam, the Netherlands.
E-mail address: Elmarije.vanStraalen@gmail.com (E.K. van Straalen).

[1] According to the Oxford Dictionary, expertise is “expert knowledge or skill in a particular subject, activity or job”.

https://doi.org/10.1016/j.fsisyn.2023.100437
Received 24 January 2021; Received in revised form 9 September 2023; Accepted 10 September 2023
Available online 29 September 2023

material to investigate the hypothesis that trace and reference material come from the same source. The lay-out and language use in these forensic reports differs greatly not only depending on the institutions involved, but also between fields of expertise and even between individuals [3,34]. In practice, conclusions stating the results of this comparison and its evidential value are reported in different ways as well. In this paper we will focus on numerical likelihood ratio (NLR), verbal likelihood ratio (VLR) and verbal categorical (CAT) conclusions. The NLR conclusion is used when the evidential value can be calculated statistically (e.g., DNA). If a statistical calculation is not possible (e.g., shoemarks), the evidential value can either be expressed in verbal terms coming from an ascending scale (i.e. weak, moderate, moderately strong, strong, etc.) or categorical conclusion (i.e. individualisation). Research has shown that the conclusions of forensic reports can be difficult to understand for its users. The goal of this study is to find out whether these conclusions are clear for criminal justice professionals and students, and whether education and practical experience help to understand these conclusions.

Numerous studies have examined the interpretation of forensic conclusions. The results of these studies can be of relevance for daily practice and for the development of better adjusted and more comprehensible forensic reports, conclusions, instructions, and refined education plans. In general, participants in these studies are asked to assess forensic evidence (reports) with various types of conclusions: numerical LR, verbal LR, visual LR, tables, categorical, etc. Participants in these studies can be criminal justice professionals who assess forensic evidence in their daily practice, students who might have some education in assessing forensic evidence, and/or lay persons who are not familiar with assessing forensic evidence and reports. In our study we focus on students who might have some education in assessing forensic evidence, and criminal justice professionals consisting of legal professionals and crime investigators.

Lay people without any education or experience in assessing forensic evidence make a variety of misinterpretations when assessing forensic reports. While in one study strong conclusions were correctly assessed as being stronger than weak conclusions, not all strong conclusions with comparable strength were assessed as such [5]. In another study one conclusion type was assessed as being weaker than other conclusion types even though these were in fact of comparable strength [6]. When lay participants were asked to assign numerical values to verbal probabilities, these values appeared to vary widely between participants [7]. (Psychology) students as participants show different types of misinterpretations: weak conclusions were overvalued while strong conclusions were undervalued [8], and mostly low-strength evidence was misinterpreted [9].

Multiple studies on the interpretation of forensic conclusions by criminal justice professionals indicate that understanding forensic reports is also difficult for professionals. Criminal justice professionals, such as judges, criminal lawyers, public prosecutors, forensic experts, police detectives, and crime scene investigators, all misinterpret the weight of forensic conclusions to some degree [1,2,4,10]. In a previous study [4] we examined how criminal justice professionals interpreted weak and strong CAT, VLR and NLR conclusions in forensic reports. The results showed that the professionals overestimated the strength of almost all the conclusions. This was the case for all strong conclusions and for most of the weak conclusions, except for the weak CAT conclusion, which was underestimated. Overall, the professionals answered about a quarter of all the questions measuring their actual understanding of the conclusions incorrectly [4]. In addition, van Straalen et al. [4] discuss a possible difference in understanding between the uncertainty of the evidence and the strength of the evidence.

Various studies examined differences in interpretation by professionals with a legal background or a police background. Keijser et al. [2] showed that legal professionals more often made the prosecutor's fallacy[2] than crime investigation professionals. However, there was no significant difference in their understanding of different types of conclusions (VLR of visual LR) [1,2]. In the study by Arscott, Morgan, Meakin, and French [11], legal professionals assessed a verbal scale in the same way as forensic professionals and students. However, legal professionals were better at assessing the conclusion with lower evidential strength, whereas the forensic professionals and students were better at assessing the conclusion with higher evidential strength [11]. Our previous study showed that legal professionals answered more factual questions about forensic conclusions correctly than did the crime investigation professionals [4]. Furthermore, this and other studies showed that criminal justice professionals overestimate their own knowledge of the interpretation of forensic conclusions [1,2,4,10]. It can therefore be very helpful if these professionals are advised on the interpretation of forensic reports. Courts in the Netherlands have employed forensic advisors with an academic background in forensic science, who advise judges in the understanding and logical correct interpretation of forensic reports [35].

A number of researchers have compared the assessment of forensic conclusions by experts and novice/lay participants. Some did not show any significant differences in the assessment of forensic conclusions by experts and novices [11,12]. Others did find differences, such as McQuiston-Surrett and Saks [13]. They examined how variations in the presentation of the forensic evidence affected judgements in trial. The study showed that professionals were not influenced by the type of conclusion that was used to present the evidence while the novices were.

1.2. Becoming an expert and the influence of expertise on decision making

There are several factors affecting decision making, in relation to this paper we focus on the influence of experience. In this section we will discuss becoming an expert, the role of feedback and the influence of experience on decision making.

How does someone become an expert? One of the first studies on this topic was by De Groot [14]. In a series of experiments, he presented chess positions to players with different levels of experience. Each player was asked to make the next move on the board and to express each thought aloud. Both experts and novices seemed to apply the same decision-making steps. However, the experts solved problems faster, more precisely and more efficiently than novices by using an automatic thinking process. De Groot identified two aspects of memory, namely 'knowledge' and 'intuitive experience' and found that, while knowledge can be explained and passed on by an expert, intuitive experience can only be explained and passed on when the expert becomes aware of this experience. However, experts are often unaware of this experience, and only use it automatically when necessary [14]. This intuition seems to influence various types of experts, including medical doctors [15]. However, this unconscious experience is problematic when it is based on incorrect assumptions or thought processes, and can lead to individuals intuitively making the same mistakes over and over again. Especially in situations where such errors or inefficiencies are not noticed due to lack of feedback on or consequences of these actions. Dreyfus and Dreyfus [16] identified five stages in the process of becoming an expert, namely novice, advanced beginner, competent, proficient, and expert. A *novice* follows a set of 'context free' learned rules. When attaining more experience and learning more rules per situation, one will recognize the context in which the rules are applied and becomes an *advanced beginner*. To become *competent,* one needs to be able to organize rules in a hierarchical structure with rational perspective of which rules are applied in which situation. The *proficient* stage is a transition between being competent and becoming an expert. An *expert* does not analyze a set of

[2] See Section 1.1 in Van Straalen et al. [4] for an explanation on the prosecutor's fallacy.

rules but goes directly to the solution [16]. For someone to truly evolve into an expert, practice in complex problem solving in the relevant field is necessary [16]. The required deliberate practice is a gradual process in which suitable and challenging training tasks are designed by a coach or teacher, which need much practice, concentration and feedback to be mastered [17]. In the different stages of becoming an expert, someone attains the characteristics common to an expert. Experts have more and better organised and integrated knowledge, and notice important patterns of information. Their strategies for accessing and using knowledge are better than those of novices, and experts are more self-regulated [18–20].

Multiple studies examined the effect of feedback on performance. Results of the study by Lichtenstein and Fischhoff [25] showed that providing feedback improved participants' ability to correctly perform tasks. Those who already performed well before the feedback showed the least improvement. There are several studies examining the effect of cognitive feedback components: task information, cognitive information, and functional validity information [26–28]. In all studies, the performance of participants improved significantly after receiving feedback with information about the task they were asked to perform, compared to receiving no feedback or feedback containing cognitive or functional validity information. However, the participants in these studies were professionals who were used to receiving direct feedback about their decisions. For example, a chess player receives feedback from his opponent, and a doctor receives feedback from her patients. This is not the case for criminal justice professionals assessing forensic reports, since they do not usually receive feedback on their assessments of these reports.

What is the influence of experience on decision making? Studies in various fields of expertise have shown that experts are faster, more accurate, notice more information and potential solutions, think beyond the problem, and approach problems on a more abstract meta-level than do novices [21–24]. Studies on the influence of experience on decision making in the field of forensic evidence show a nuanced picture. Van den Eeden et al. [29] studied the influence of context information on the forensic decisions of theoretically educated forensic science students and experienced crime scene investigators. Their results showed no difference in the number of secured traces between students and professionals. However, the students did secure more crime-related traces, and were more confident about their first impressions of the crime scene [29]. These results show that only having experience is not sufficient, and that education in this field is important as well. In a study by Baber and Butler [30], a mock crime scene was examined by expert and novice (student) crime scene examiners. While the novices were more focused on individual objects and on reconstructing what had happened, the experts were faster, and were more focused on fewer objects with more evidential value.

The above literature shows that while evolving from novice to expert, experts acquire qualities enabling them to become better at certain tasks. However, studies have shown that this does not necessarily mean that experts are better at all tasks than are novices. Repeated training tasks with immediate feedback is needed to gain expertise [17]. For someone to truly evolve into an expert, experience alone is not enough. Deliberate practice is required for the maintenance of expertise and is enhanced by feedback [17]. When there is a lack of feedback about experts' performance, the experts cannot learn from their mistakes and do not know when or how their decisions could be improved. Most forensic examiners build their experience by their casework in which the ground truth is unknown. Since their work therefore mostly cannot provide the needed feedback, proficiency tests are necessary but mostly not implemented in a way to meet the deliberate practice standards [31].

Once forensic conclusions are misinterpreted, this not only may affect the work of the professional making the mistake, but also may influence the work of every professional who bases his or her work on the mistaken interpretation. Thus, when mistakes are made in the criminal justice chain, this may cause bias to snowball [32]. Therefore, it is important to have more insight into the interpretation of forensic reports and conclusions, and the influence of education and experience on this interpretation. These insights can be used for creating awareness of a possible lack of understanding and for training purposes. This may help to avoid bias in forensic decision making throughout the criminal justice process.

1.3. *The present study*

The results from our previous study showed how criminal justice professionals make mistakes in their assessment of forensic conclusions [4]. In the current study, we compare students to these professionals both assessing the same evidence.

A typical difference between students and professionals is their level of experience. Since students lack experience, they may be less equipped to assess the evidential value of forensic conclusions. Criminal justice professionals are more experienced and often have backgrounds in judging cases and handling forensic evidence. It may be hypothesized that those with more experience are better at judging forensic reports. However, if their previous decisions were not entirely correct and they did not receive any feedback on these assessments, they may have gained experience in making inaccurate interpretations and decisions, which means they did not become experts but rather are trained in making the same mistakes repeatedly.

The aim of this study is to investigate whether education and experience influence the ability to assess forensic conclusions correctly. As mentioned in paragraph 1.1, there might be a difference in the understanding of the uncertainty of evidence and the strength of evidence in forensic conclusions. This has not explicitly been analysed in the studies mentioned earlier and will be included in the current study. We will attempt to answer the following questions: Do students and criminal justice professionals misinterpret forensic conclusions? Are criminal justice professionals better at recognising the uncertainty and strength of evidence while assessing different types of forensic conclusions compared to students? Does having a legal or crime investigation educational background influence the student's ability to assess forensic reports, and does this ability improve with gaining professional experience in these fields? Are criminal justice professionals better at assessing their own understanding of forensic conclusions compared to students? Based on these questions and the literature, we have three main hypotheses.

Firstly, we expect there to be a significant difference in the assessment of forensic conclusions between students and professionals. Based on research into the influence of experience on professional decision making, we expect that professionals in the criminal justice system are better able to recognize relevant information in forensic reports and conclusions than students. To train this competency, experience might not be enough. Developing expertise requires more than just formal education and experience. Appropriate and immediate feedback seems to be crucial in this process [31]. As feedback is often missing in the forensic context, criminal justice professionals may not perform much better than novices. Students are in the process of receiving theoretical education on the value of forensic reports. We presume this provides them with the necessary knowledge to assess forensic conclusions correctly. However, students lack experience in reading and assessing forensic conclusions in case work. Although gaining experience does not necessarily lead to expertise, we do think that experience offers added value over just theoretical knowledge. Therefore, we think criminal justice professionals will outperform students.

Secondly, based on research on the assessment of forensic evidence by criminal justice professionals, we expect both students and professionals with a legal background to be better at assessing forensic conclusions than students and professionals with a background in crime investigation.

Thirdly, we expect to see a higher self-proclaimed understanding for criminal justice professionals compared to students. Since students lack experience in assessing forensic reports, we assume them to be less confident of their ability to correctly assess forensic reports and conclusions than criminal justice professionals.

The results of our study can be of help for both the training for students and criminal justice professionals, and for adjusting the format and content of forensic reports and conclusions.

2. Method

2.1. Design

This article describes the analyses of a combination of two separate studies using the same method but a different participant population: criminal justice professionals were included in study II, and students learning to become criminal justice professionals in study I. The results of study II are presented in Van Straalen et al. [4]. The present study focusses on similarities and differences between the two participant groups. The analyses will be explained in Section 2.4.

In both study I and study II, the participants took part in an experiment that entailed an online questionnaire that presented reports of a fingerprint examination. The reports consisted of a simplified one-page summary of a police fingerprint examination containing some basic fictional personal information about a suspect, identification numbers of the traces and reference material, and the conclusions of the examination. Other information about the trace, the forensic investigation methods and the actual comparison was left out. The reports were all identical except for the conclusion part, which varied in this study. In the conclusion part of each report, one of three types of conclusions was used: a categorical conclusion, a conclusion using a verbal LR, or a conclusion using a numerical LR. The reports pertained to either strong evidence or weak evidence; thus for all three conclusion types, there was phrasing expressing strong evidential strength and phrasing expressing weak evidential strength. For the weak evidential strength, the different conclusion types resulted in the phrasings of categorical—'cannot rule out', verbal LR—'moderate', and numerical LR—'LR of 50'. For the strong evidential strength conclusions this resulted in categorical—'individualisation', verbal LR—'extremely strong' and numerical LR—'LR of 5 million'. All three strong evidential strength phrasings were of comparable strength. In practice, these phrasings are used to refer to the same evidential strength. The same was true for the weak evidential strength phrasings. These three phrasings were also of comparable strength. Since for VLR and NLR conclusions an ascending scale is used, it was clear which of these conclusions are of comparable strength. Categorical conclusions do not use a scale. If there are no differences and enough similarities between a trace and comparison material, the conclusion used will be an 'individualisation' or 'cannot rule out' (depending on the number of similarities). Weak evidence that is phrased as 'moderate' or 'LR of 50' in a likelihood conclusion, is phrased as 'cannot rule out' in a categorical conclusion. This categorical phrasing 'cannot rule out' is used for a range of forensic conclusions. An often-mentioned problem with categorical conclusions is that it exaggerates evidential strength because it suppresses any notion of uncertainty [31]. However, when it comes to weak evidence, categorical conclusions offer less certainty than likelihood ratios, as there is only one phrasing that indicates a range of findings. This may lead to an underestimation of the evidence since the phrasing used does not provide a specific value.

Fig. 1 shows an overview of the survey design for studies I and II. Both student and professional participants received three reports: one report with a categorical (CAT) conclusion, one report with a verbal LR (VLR) conclusion, and one report with a numerical LR (NLR) conclusion. The evidential strength of the conclusions could be strong or weak. All three of the reports that the students received could contain only conclusions with weak evidential strength, only conclusions with strong evidential strength, or a combination of conclusions with strong and weak evidential strength. After study I, we decided to simplify this design because the difference in presenting 'weak *and* strong' conclusions or 'weak *or* strong' conclusions did not influence the results. Therefore, all three reports that the professionals received could contain either only conclusions with weak evidential strength, or only conclusions with strong evidential strength. The students were randomly allocated to one of eight conditions that differed in the order of the reports being presented. The professionals were randomly allocated to one of six conditions. The difference in number of conditions was due to the slight change in design, as explained above.

2.2. Participants

Participant groups were selected from people who, in their (future) jobs, may be assessing forensic reports and conclusions. In study I, 96 students[3] took part in the questionnaire. The student population consisted of students of Forensic Science and Forensic Investigation[4] (45%), Law (39%), Criminology (8%), and Detective Education at the police academy (7%). In study II, 269 professionals took part in the questionnaire. The professionals[5] consisted of crime scene investigators (23%), police detectives (28%), public prosecutors (21%), criminal lawyers (11%), and judges (17%). Table 1 shows the different characteristics of the students, the professionals, and the total group of participants. Of the total participant population, 49% was female. There were more females in the student participant population (71%) compared to the professional participant population (43%). The mean age of the total participant group was 41 years (SD = 13). In general, the professionals were older (M = 45) than the students (M = 25), which was to be expected.

2.3. Procedure

For study I, the participants were recruited via their teachers, and through specific student pages on the online social media website Facebook. For study II, the participants were recruited via an email that was sent out within their organisations. Participation was voluntary for both participant groups. Students could include their email addresses at the end of the questionnaire to have the opportunity to win a gift card of €25. The recruitment text for both participant groups contained an URL directing them to the online questionnaire on the survey website www.qualtrics.com. In both studies, first a short welcome text appeared, explaining that the study focused on the interpretation of forensic reports, and that three fingerprint examination reports would be presented with corresponding questions. For study I, the first question after the welcome text inquired about the participant's type of education. For study II, the welcome text was followed by several questions about the participant's current job position and other background characteristics. The next questions in both studies were identical. A forensic report with a set of questions was presented three times in succession. These questions are similar to those used in other studies [1,2,10]. For each report, six questions measured the alleged understanding of the report. These were open text and five-point Likert scale questions. Of these six, the three questions using a five-point Likert scale were eventually used in the analyses. See Section 3.1 for a detailed description of the questions measuring alleged understanding. Eleven questions measured the actual understanding of the conclusion, the guilt of the suspect and the level of incrimination of the report. These questions had answer options 'yes', 'maybe', 'no', 'don't know' (Q1 and Q3); 'yes', 'no', 'don't know' (Q2, Q4, Q5, Q6, Q7, Q8, Q9); or five-point Likert scales (Q10 and Q11). For

[3] When the term 'student' is used, this means all students in this study.

[4] A practice-oriented bachelor's and master's education for future professionals in the forensic field.

[5] When the term 'professional' is used, this means all professionals in this study.

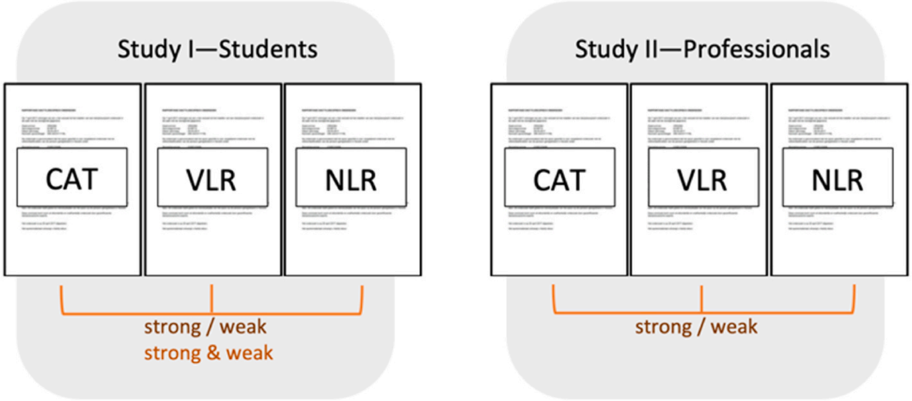


Fig. 1. Survey design for Students in Study I and Professionals in Study II.

Table 1
Participant characteristics.

	N	Age (M)	Gender (% female/male)
Students	96	25 (SD = 8)	71/29
Professionals	269	45 (SD = 11)	43/57
Total	365	41 (SD = 13)	49/51

the analyses, the answers to Q1-Q9 were divided into ‘correct’ and ‘incorrect’, based on the classification used in other studies [1,2,10]. See the tables in Section 3 for a detailed description of the questions measuring actual understanding. If a participant assessed a report and answered the corresponding questions, he or she could move on to the next report and corresponding questions. Once having moved on, the participant could not return to the previous page. After assessing three reports and answering all the questions, a final set of questions appeared to measure the participants’ experience with fingerprint examination reports and forensic reports in general. In both studies, the participants were questioned about their background characteristics. In study II, the participants were also asked about several procedures concerning fingerprint evidence (reports) within their organisation. At the end of the questionnaire, the participants could leave a comment about the questionnaire, or about forensic (fingerprint) evidence reports in general.

2.4. Data analysis

All data were exported from the survey website www.qualtrics.com and analysed. First, we analysed whether the difference in the designs of study I (students) and study II (professionals) influenced the results. For these analyses, the total population was divided into six groups:

1) students assessing three strong conclusions; 2) students assessing two strong and one weak conclusion; 3) students assessing one strong and two weak conclusions; 4) students assessing three weak conclusions; 5) professionals assessing three strong conclusions; and 6) professionals assessing three weak conclusions. The groups of students with at least one strong conclusion (groups 1, 2 and 3) were compared to the group of professionals with strong conclusions (group 5). The groups of students with at least one weak conclusion (groups 2, 3 and 4) were compared to the group of professionals with weak conclusions (group 6). The numbers of correct answers to the questions measuring actual understanding for the strong or weak reports were compared within these two sets of groups.

In study II [4], the results of the professionals were analysed on the report and question level; that is, the percentage of assessed reports with a correct answer per question. Additional analyses were performed in the current study. The answers to all nine questions measuring the actual understanding were first divided into correct = 1 and incorrect = 0. Next, the mean total amount of correct answers to those nine questions was calculated per report, per person (total of all three reports), per evidential strength and per type of conclusion used in the report. We studied the influence of having a legal or crime investigation background. Therefore, we divided the group of participants into a) students with a crime investigation background: students of Forensic Science, Forensic Investigation, Detective Education at the police academy, and Criminology; b) students with a legal background: students of law; c) professionals with a crime investigation background: crime scene investigators and police detectives; and d) professionals with a legal background: public prosecutors, criminal lawyers, and judges.

In total, the following three analyses were performed for conclusion type, evidential strength, experience (students vs. professionals), and legal or crime investigating background: 1) on the *question* level – the mean percentage of reports with a correct answer to each question measuring actual understanding; 2) on the *report* level – the mean amount (ranging from 1 to 9) of correct answers to the nine questions measuring actual understanding; 3) on the *question* level – the mean answers (ranging from 1 to 5) to the three questions measuring self-proclaimed understanding.

At all levels of analyses, the effects of the evidential strength (strong, weak) and of conclusion type (NLR, VLR, CAT) of the report were measured using a one-way ANOVA, and the effect of type of participant assessing the report was measured using a two-sample *t* test.

3. Results

The main goal of this study was to examine whether criminal justice professionals are better at assessing forensic conclusions compared to students lacking professional experience. Firstly, we analysed the difference in the designs for the student and professional groups, meaning that students could receive reports with strong *and*/or with weak conclusions, whereas professionals could only receive reports with strong *or* with weak conclusions. A one-way ANOVA showed no significant difference between the groups of students and professionals assessing at least one report with a strong conclusion (F (3,495) = 0.154, p = .93), and no difference between the groups of students and professionals assessing at least one report with a weak conclusion (F (3,494) = 0.338, p = .80). Therefore, it can be concluded that the design had no effect.

In this section, the results are presented concerning the understanding of the forensic conclusions. The first results presented pertain to the effect of evidential strength and conclusion type on the understanding and assessment of forensic conclusions for the total group of participants, followed by results comparing the different participant groups. Secondly, we consider the effect of having experience with fingerprint evidence. Lastly, we look at the self-proclaimed understanding of the participants.

3.1. *Actual understanding*

We analysed the actual understanding of the reports and conclusions on two different levels:

1) on the *question level*, the mean percentage (0–100%) of reports with a correct answer, and on *question set level* measuring either the *uncertainty* or the *strength of evidence* (Tables 2 and 3),
2) on the *report level*, the mean number of correct answers (0–9) to the nine questions measuring actual understanding (Table 4).

3.1.1. *Actual understanding –general findings on question level, uncertainty, and strength*

The general understanding of the conclusions and evidential strengths was measured using nine questions about the (level of) incrimination of the evidence, and the guilt of the suspect. To obtain a more detailed insight into the understanding of the conclusions, for these nine questions the percentage of reports with correct answers was analysed. Since in general the students and professionals were quite similar in their understanding of the conclusions, we will first describe the understanding of the total group of participants. Next, we will describe differences between the background and the experience of the participants. Table 2 presents the percentage of correct answers per question for the total group of participants divided into evidential strength and type of conclusion.

Looking at the total group of participants for all questions there was a significant effect of the evidential strength on the percentage of correct answers (see Table 2). For 6 out of 9 questions most correct answers were provided for reports with weak evidential strength. For all nine questions, the conclusion type for these reports had a significant effect on the percentage of correct answers. This seemed to be caused mainly by the understanding of the weak CAT conclusion compared to the other weak conclusion types. For 6 of the 9 questions the weak CAT conclusion was understood best, for the other 3 questions the CAT conclusion was understood poorer than the other weak conclusion types. For the strong evidential strength reports, this effect of conclusion type was only found for 2 out of 9 questions. For these two questions this was caused by the (poorer) understanding of the CAT conclusion compared to the other strong conclusion types. The sole question containing a prosecutor's fallacy (Q9), "There is more than a 50% chance the fingerprint belongs to the suspect", was answered inaccurately most often. Participants gave a correct answer to this question in only 27% of all the assessed reports. Especially for reports with a strong conclusion, this question was answered incorrectly (91%). Participants more often answered this question correctly for reports with a weak conclusion, and particularly for reports with a weak CAT conclusion.

Since the questions we asked either measured the understanding of the *strength of the evidence* or the understanding of the *uncertainty of the evidence*, all the questions measuring actual understanding were allocated to one of these two groups. Except for Q9 which we discussed before. This enabled us to measure whether the *uncertainty* (Q1, Q2, Q3, Q4, Q5) and/or the *strength of evidence* (Q6, Q7, Q8) were assessed more accurately for certain conclusion types and by certain types of participants. The mean answers for these two sets of questions are presented in Table 2 in blue. For both sets, the evidential strength of the conclusion presented in the report had a significant effect on the percentage of correct answers. Questions about the *strength of the evidence* were answered correctly more often for the strong conclusions than for the weak conclusions. Questions about *uncertainty* were answered correctly more often for the weak conclusion types. For the strong conclusion types, only the answers to the questions about *uncertainty* differed significantly depending on the phrasing of the conclusion (F (2,496) = 3.844, p < .02). Participants assessed the strong CAT conclusion more often (incorrectly) as providing 100% certainty compared to the other strong conclusion types. For the weak conclusion types, the answers for both groups of questions on *uncertainty* (F (2,495) = 17.513, p < .001) and *strength of evidence* (F (2,507) = 214.086, p < .001) differed significantly by conclusion type. Questions about the *strength of evidence* were answered significantly less often correctly for the weak CAT conclusion compared to all the other conclusions.

Table 3 presents the percentage of correct answers per question for students and professionals and their different backgrounds. Although there were differences between the students and professionals in their understanding of the *strength of the evidence* and the *uncertainty*, the only significant difference was in the understanding of *uncertainty* by professionals: L professionals answered significantly more questions correctly (M = 0.82, SD = 0.23) than did CI professionals (M = 0.66, SD = 0.32), (t (740) = 7.82, p = .001).

When we look at question level, in general the students and professionals had an equal understanding of the conclusions. Nonetheless, there were two questions for which there was a significant difference in their understanding. For the question 'It has been proven that the suspect was at the scene where the finger mark was found. (no)', the

Table 2
The percentage of correct answers per type of conclusion and evidential strength.

Questions and statements	Mean for all reports	*Conclusion strength and type*							
		Weak				Strong			
		Total weak	CAT (A)	VLR (B)	NLR (C)	Total strong	CAT (D)	VLR (E)	NLR (F)
1. Does the finger mark belong to the suspect? (maybe)	57*	77**	93	68	70	**38**	34	41	39
2. Do you think it is impossible for the finger mark to be from someone other than the suspect? (no)	73*	83**	91	78	81	63**	54	68	66
3. It has been proven that the defendant is guilty. (no)	89*	93**	97	94	90	**84**	84	84	83
4. It has been proven that the suspect was at the scene where the finger mark was found. (no)	63*	71**	87	65	62	**55**	52	57	56
5. It is ruled out that the finger mark belongs to someone other than the suspect. (no)	84*	88**	95	84	85	80**	69	84	87
Questions about uncertainty	73[a]	83[c]	92	78	78	64[c]	59	67	66
6. The outcome of this examination is evidence against the suspect. (yes)	70*	57**	25	68	79	**83**	82	84	84
7. The result of this examination is incriminating for the suspect. (yes)	72*	62**	25	75	87	**83**	81	84	84
8. The conclusion better fits the scenario that the finger mark belongs to the suspect than the scenario that it belongs to someone else. (yes)	80*	65**	17	87	92	**95**	94	95	95
Questions about strength of evidence	74[b]	61[c]	22	77	86	87	86	88	88
9. There is more than a 50% chance the finger mark belongs to the suspect. (no)	27*	45**	60	31	42	**9**	7	8	11

Note: *Significant effect of evidential strength on total at p < .001 level. ** Significant effect of conclusion type on total weak or total strong at p < .001 level. [a] Significant effect of evidential strength on mean correct answers (F(1,995) = 114.970, p < .001). [b] Significant effect of evidential strength on mean correct answers (F (1,1014) = 141.919, p < .001). [c] Significant effect of conclusion type on mean answers for all weak or all strong options at p < .001 level.

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

Questions and statements	Mean for all reports	*Experience and background*					
		Student			*Professional*		
		Total student	Legal	Crime Investigation	Total prof.	Legal	Crime Investigation
1. Does the finger mark belong to the suspect? (maybe)	57	**58**	55	60	***57****	65	49
2. Do you think it is impossible for the finger mark to be from someone other than the suspect? (no)	*73*	***71***	*68*	*73*	***74****	*86*	*62*
3. It has been proven that the defendant is guilty. (no)	*89*	***88***	*85*	*91*	***89***	*91*	*87*
4. It has been proven that the suspect was at the scene where the finger mark was found. (no)	*63*[†]	***58****	*49*	*65*	***65****	*75*	*56*
5. It is ruled out that the finger mark belongs to someone other than the suspect. (no)	*84*	***85***	*87*	*84*	***84****	*93*	*75*
Questions about uncertainty	*73*[a]	***72***	*69*	*74*	***74***[b]	*82*	*66*
6. The outcome of this examination is evidence against the suspect. (yes)	*70*[†]***	***72***	*69*	*74*	***70***	*72*	*67*
7. The result of this examination is incriminating for the suspect. (yes)	*72*	***74***	*76*	*73*	***72***	*69*	*75*
8. The conclusion better fits the scenario that the finger mark belongs to the suspect than the scenario that it belongs to someone else. (yes)	*80***	***80***	*83*	*78*	***80***	*79*	*80*
Questions about strength of evidence	*74*[c]	***75***	*76*	*75*	***74***	*73*	*74*
9. There is more than a 50% chance the finger mark belongs to the suspect. (no)	*27*	***29***	*26*	*31*	***26****	*30*	*22*

Note: [a] Significant effect of evidential strength on mean correct answers (F(1,995) = 114.970, p < .001). [b] Significant effect of participant background (legal or crime investigation) on mean correct answers (t(740) = 7.82, p = .001). [c] Significant effect of evidential strength on mean correct answers (F(1,1014) = 141.919, p < .001). [†] Significant effect of 'experience' (student or professional) at p < .05 level. *Significant effect of legal background for students or professionals at p < .001 level. **For the total group of participants, the significant effect of being a legal or crime investigation participant on the percentage of correct answers per question at p < .001 level.

professionals overall answered this question more often correctly (M = 0.65, SD = 0.478) than did the students (M = 0.58, SD = 0.494), (t (1014) = −1.888, p = .001). Especially for the reports with weak evidential strength conclusions, this question was answered mostly correct by the professionals (M = 0.73, SD = 0.447) compared to the students (M = 0.68, SD = 0.470), (t (508) = −1.072, p = .045). When we look more closely (Table 3), we see that most correct answers for this question were given by the legal professionals. For reports with strong conclusions, the question 'The outcome of this examination is evidence against the suspect. (yes)' was significantly more often correctly answered by the students (M = 0.88, SD = 0.332) compared to the professionals (M = 0.81, SD = 0.391), (t (504) = 1,629, p = .001).

Looking more closely at participants backgrounds (legal (L) vs. crime investigation (CI)), we see that CI students generally outperform L students, and that L professionals generally outperform CI professionals. For both students and professionals this difference was reversed for questions about the level of incrimination of the evidence (Q7 and Q8).

3.1.2. Actual understanding – report level

To obtain an overview of how well the reports were assessed in general, for the nine questions measuring the actual understanding the number of correctly answered questions were added up per report. Table 4 shows the average number of correct answers per report, for conclusion type and evidential strength for L professionals and students, and CI professionals and students. In general, the average number of correct answers for all the assessed reports and all the participants was 6.2 out of 9. Reports with weak evidential strength conclusions were assessed significantly better than reports with strong evidential strength conclusions. In general, for all participants and all the reports, the number of correct answers per report differed significantly according to the type of conclusion in the report. In particular, the weak and strong CAT conclusions stood out, and the understanding of this conclusion type was poor compared to that of the VLR and NLR conclusions. Looking at participants' background, students and professionals were similar in their assessment of the reports, their conclusions, and their evidential strengths. The only significant different was found for the

Table 4
Average number of correct answers per report out of nine questions about actual understanding.

Evidential Strength	Conclusion type	Experience (student vs. professional), Background (L vs. CI)						Total
		Students			Professionals			
		Total stud.	Legal	Crime investigation	Total prof.	Legal	Crime Investigation	
Strong	*Categorical**	***5.7***	5.5	5.8	***5.6***	6.3	5.0	**5.6**
	*Verbal LR**	***6.1***	5.9	6.2	***6.0***	6.5	5.4	**6.0**
	*Numerical LR**	***6.0***	5.8	6.2	***5.9***	6.3	5.5	**5.9**
*Total strong**		***5.9***	***5.8***[ab]	***6.0***[ab]	***5.9***	***6.4***[a]	***5.4***[b]	***5.9*****
Weak	*Categorical**	***5.7***	5.6	5.8	***5.9***	6.1	5.7	**5.9**
	*Verbal LR**	***6.7***	6.5	7.0	***6.7***	7.3	6.2	**6.7**
	*Numerical LR**[†]	***6.6***	6.1	7.0	***7.0***	7.6	6.6	**6.9**
*Total weak**		***6.4***	***6.2***[a]	***6.6***[ab]	***6.5***	***6.9***[b]	***6.1***[a]	***6.5*****
TOTAL		***6.2***	***6.0***[ab]	***6.3***[a]	***6.2***	***6.7***	***5.8***[b]	***6.2***

Note:*Significant effect of background on M of correct answers per type of evidential strength and conclusion type for professionals at p < .001 level. [†]Significant effect of background on M of correct answers per type of evidential strength and conclusion type for students at p < .001 level.
The total means in the same row that do not share the same superscripts differ at p < .05. **Significant effect of conclusion type on average number of correct answers per evidential strength at p < .001 level (F(2,994) = 17.832, p < .001).

CAT reports. CAT reports in general were slightly better understood by professionals (M = 5.76, SD = 1.626) than by students (M = 5.70, SD = 1.385), (t (985) = −0.506, p = .017). Whereas for the strong CAT reports students slightly answered more questions correctly (M = 5.69, SD = 1.522) than did professionals (M = 5.61, SD = 1.797), (t (497) = 0.413, p = .006).

We will further focus on the effect of the background of the participants (L vs. CI) and on the interaction between background (L vs. CI) and experience (student vs. professional). Participants with a legal background answered 6.5 out of 9 questions correctly, while this was 5.9 out of 9 for CI participants. The CI students performed better than the L students for all the evidential strengths and conclusion types, and the L professionals performed better than did the CI professionals. All these differences were only statistically significant for the professionals.

Fig. 2 shows the differences between the four participant groups in terms of their interpretations of the conclusion types in more detail. It shows that, for the strong and weak conclusions, all the participant groups had poorest understanding of the CAT conclusion. While the students' assessments of the VLR conclusions were better than or comparable to their assessments of the NLR conclusions, the professionals were generally better at assessing the NLR conclusions than the VLR conclusions.

Fig. 3 shows the interaction for the effect of background (legal vs. crime investigation) and experience (student vs. professional) on the correct assessment of forensic conclusions.

3.2. *Experience with fingerprint/forensic evidence*

To answer the question if experience helps in assessing forensic reports, we compared different groups of students and professionals. Because these groups were quite broad, we also collected specific information about the education and experience of individual participants concerning the assessment of forensic fingerprint reports (see Table 5). These additional questions were presented at the end of the questionnaire, and were answered by 238 of the 269 professionals, and by 73 of the 96 students.

Participants were asked whether they had ever read a fingerprint examination report from the police, using a CAT conclusion or from the Netherlands Forensic Institute (NFI),[6] using a VLR or NLR conclusion prior to this study. None of the L students had any experience with these reports. About half of the CI students had some experience with police or NFI reports on fingerprint evidence. Of the professionals, a large number were familiar with assessing forensic reports. The L professionals had more experience with fingerprint reports of the NFI, using VLR or NLR conclusions whereas the CI professionals had more experience with fingerprint reports from the police, using CAT conclusions. Just over half of the professionals and students showed to have basic knowledge on the interpretation of forensic fingerprint conclusions. They correctly indicated that fingerprint examiners cannot be 100% certain about the source of a fingerprint. More L professionals than CI professionals gave a correct answer to this question. For the students the reverse was true: more CI students than L students answered this question correctly. Most of the professionals have had education about fingerprint evidence. Of them, 44% had gained theoretical knowledge by taking a course, 34% by reading (scientific) literature, and 18% through a symposium or conference. Most CI students had read literature on fingerprint evidence, whereas most L students had not.

It was found that having experience with police or NFI reports did not influence the number of correct answers for the total group of participants. Moreover, the theoretical knowledge they obtained through courses, literature and conferences did not have an effect on their assessment of the conclusions either. The only effect we observed was a strong correlation between the basic understanding that fingerprint examiners cannot be 100% certain about their conclusions regarding the source of a fingerprint and the percentage of correctly answered questions. The participants who answered this question correctly performed significantly better on questions measuring their actual understanding of forensic conclusions than those who did not (M = 19.3 vs. M = 17.4; F (1.892) = 40.769, p < .001).

3.3. *Self-proclaimed understanding – question level*

Participants were asked three questions measuring their self-proclaimed understanding:

Q1 – *Do you understand the conclusion in the report?*
Q2 – *Do you generally understand the content of the report?*
Q3 – *Do you think there is sufficient information in the report to understand the conclusion?*

For all three questions, an answering scale ranging from 1 -not at all-, to 5 -completely- was used. The answers to the questions concerning alleged understanding were analysed on the report level. Table 6 presents the mean alleged understanding on the report level per type of conclusion and evidential strength for the total group of participants. In general, the participants thought they understood the conclusion and content of the reports well. They considered there to be sufficient information in the report for them to understand the conclusion. There was an effect of evidential strength: Participants assessing reports with strong evidential strength were significantly more positive about their understanding of these reports and the conclusions than were those assessing reports with weak evidential strength and conclusions. All three questions measuring self-proclaimed understanding were answered similarly for all three conclusion types; the participants did not seem to think they understood one or two types of conclusions better than the other(s).

There was no significant difference between students and professionals in their alleged understanding. The only significant differences we saw were between legal (L) participants and crime investigation (CI) participants. The CI participants were somewhat more optimistic about the extent to which the information in the report was sufficient for understanding the conclusion (Q3: M = 3.5, SD = 1.14) compared to the L participants (Q3: M = 3.3, SD = 1.12), (*t* (943) = −2.688, *p* = 0,05). When L participants and CI participants within the student group were compared, CI students had a significantly higher alleged understanding of the conclusion and report than did the L students (Q1: legal M = 3.8, SD = 1.02; non-legal M = 4.2, SD = 0.848; F (1,246) = 9.992, p < .002), Q2: legal M = 3.8, SD = 0.872; non-legal M = 4.3, SD = 0.759; F (1,246) = 21.533, p < .001). For the professionals, there was no significant difference in alleged understanding between L and CI professionals.

4. Conclusion and discussion

The main research question in this study was whether education and experience influence the ability to assess forensic conclusions correctly.

Firstly, we examined whether professionals in the criminal justice system are better at assessing forensic conclusions correctly compared to students learning to become criminal justice professionals. Surprisingly, in general both students and professionals had a similar understanding of the forensic conclusions provided to them. Both students and

[6] The NFI has thus far produced fingerprint examination reports in a select number of cases. Considering the relatively high number of participants who reported having seen an NFI fingerprint report previously, we cannot be certain whether the participants understood this question correctly. Some participants might have thought the question was about any forensic report from the NFI. Nevertheless, we decided to analyze this question since it provides an indication of the experience of participants with reports using VLR or NLR conclusions. We expected the percentage of participants who had seen any forensic report from the NFI using a VLR or NLR conclusion to be higher than reported here.

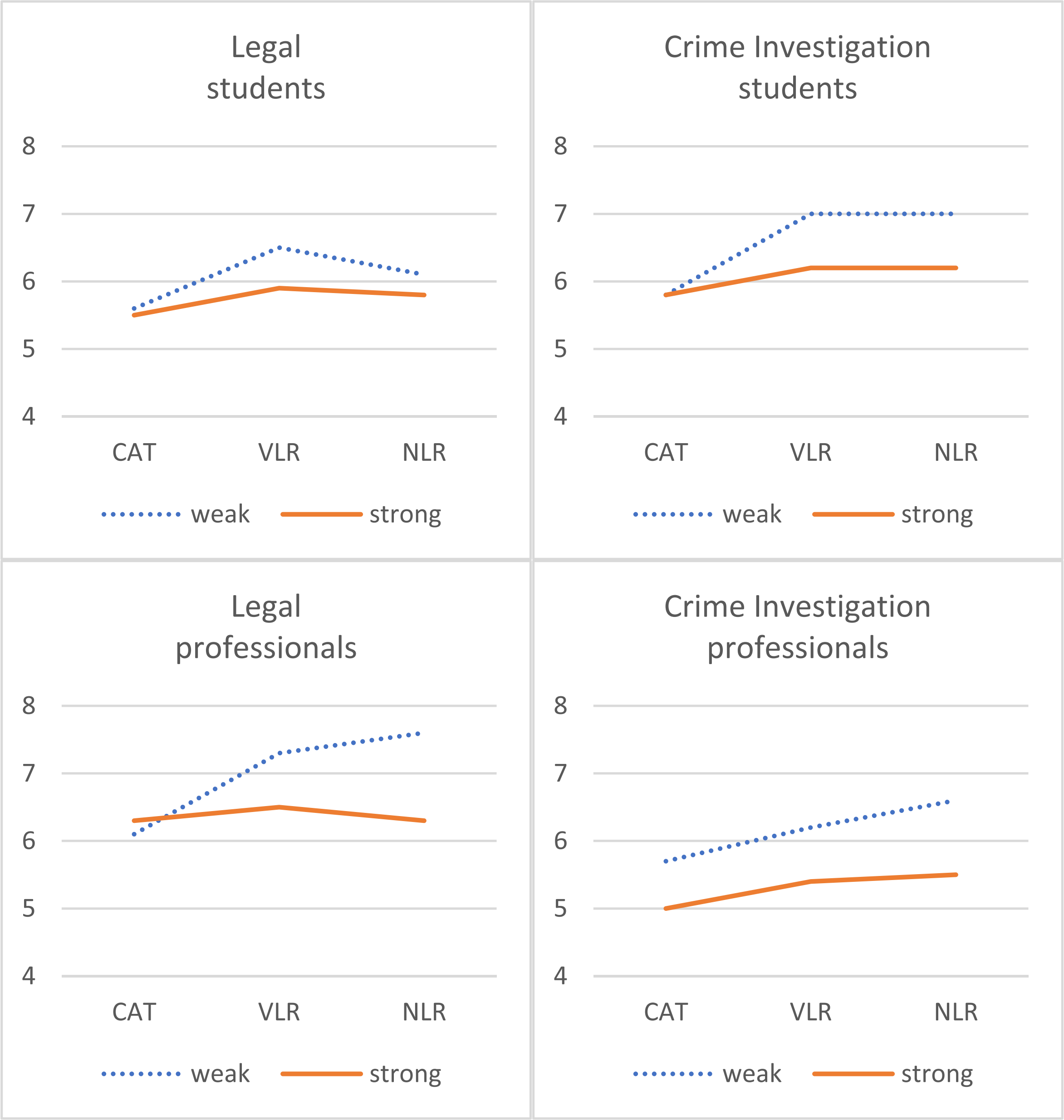


Fig. 2. Average number of correct answers (out of nine questions) about actual understanding per report, per conclusion type and per conclusion strength.

professionals had some difficulty understanding forensic reports and made mistakes. In general, they correctly answered about 75% of the questions measuring their actual understanding of the forensic conclusions. Both groups were better at assessing the correct *strength of the evidence* for the strong conclusions, especially the strong VLR and NLR conclusions, than for weak conclusions. The *uncertainty of the evidence* was understood best for the weak conclusions by both groups, particularly for the weak CAT conclusion. However, participants seemed to highly underestimate the value of weak CAT conclusions compared to the weak VLR and weak NLR conclusions. Questions about the level of *uncertainty* of the evidence were more often answered correctly for the weak CAT conclusion than for the other conclusion types, meaning that participants were better able to recognize the *uncertainty* in this type of conclusion. The *uncertainty* in the strong CAT conclusion was most often misinterpreted. When the conclusions are strong, participants seem to be able to assess the *strength of the evidence* correctly, but do not realise that there is still some *uncertainty*. For the weak conclusions, participants do not always assess the *strength of the evidence* correctly, but are better at recognising the *uncertainty* in the evidence. The only difference we found, was that professionals assessed the *strength of evidence* of the weak conclusions as being higher than did the students.

Secondly, we hypothesized that having a legal or CI background would influence the assessments of forensic conclusions. As we already knew from previous analyses, L professionals were better at assessing forensic conclusions than were CI professionals. L professionals seemed to be more hesitant about assigning too much *evidential strength* to conclusions and were better at assessing the *uncertainty* of conclusions than were CI professionals. Contrary to our expectations, L students were poorer at assessing forensic conclusions correctly compared to CI students. Overall, L students performed poorer than L professionals, and CI students performed better than CI professionals. Of the four groups, the L professionals performed best, while the CI professionals had the most difficulty in correctly assessing forensic conclusions. CI students and L professionals seemed to be more careful when assigning a certain level of incrimination to the evidence. Too careful it seems, since if evidence provides some degree of certainty about the level of

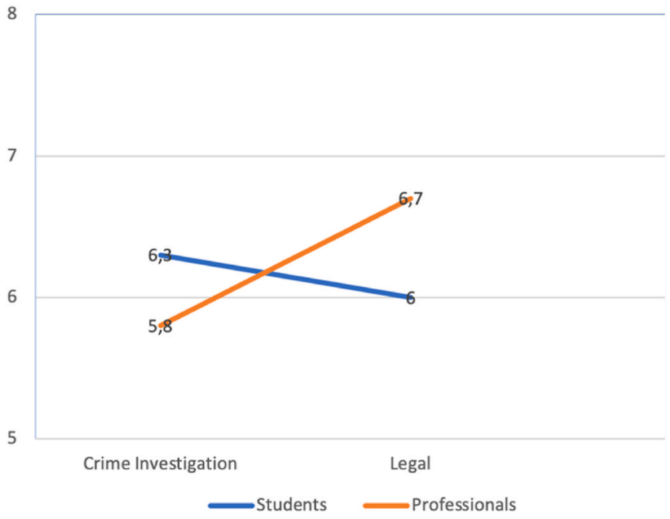


Fig. 3. Average number of correct answers (out of nine questions) about actual understanding per report.

correspondence between a trace and the reference material of a suspect, and does not rule out the suspect based on observed differences between the two, it is evidence about the probability that the suspect may be the source of the trace.

Thirdly, we expected professionals to have a higher self-proclaimed understanding of forensic conclusions than students. In general, both students and professionals thought they understood the forensic reports and conclusions we presented to them better than they actually did. Both groups felt more confident when assessing reports containing very strong conclusions or conclusions they interpreted as very weak. When they interpreted the conclusion not as strong or weak, they seemed to have more doubts about their ability to assess the strength of the conclusion correctly. This is in line with the findings in the study by Willems, Albers and Smeets [33], in which participants were better at recognising the evidential strength of more extreme values. The self-proclaimed understanding was higher for the CI students than for the L students. This corresponds to the actual understanding of these two groups. The largest group of CI students were forensic investigation students, who receive theoretical and practical courses on forensic evidence. Law students only receive basic theoretical education on forensic evidence. Besides that, the results of our study show that CI students often had some basic experience in reading forensic reports, which L students lacked. Therefore, it is not surprising that future crime investigators were more familiar with the kind of conclusions used in this study, and are more positive about their alleged understanding of these conclusions than L students. Strikingly enough however, the assumed understanding of CI professionals also turned out to be higher than that of L professionals, while in fact L professionals had a higher actual understanding of the conclusions compared to CI professionals.

To return to our main question, in general, experience did not seem to influence the assessment of forensic conclusions. However, there is an interesting interaction between profession and experience when it comes to understanding forensic conclusions: Experienced L professionals performed better than the L students, whereas experienced CI professionals performed worse than CI students. In our view, there are several explanations for this interaction.

Firstly, there is a difference in the courses these professionals take during their professional careers pertaining to the interpretation of forensic evidence. Crime investigators usually only take courses at the start of their careers or to become an expert in a new domain. Legal professionals (can) take courses throughout their careers on various topics, including forensic evidence interpretation. They receive feedback during their education, which enables them to become aware of the possible mistakes and to adjust their assessments when necessary. Without these courses and feedback, professionals may not become aware of making mistakes in the assessment of forensic evidence. Secondly, a closer look at the educational backgrounds of our participants revealed that there is a difference between the education CI professionals received in the past, and the education of current CI students. In recent years, education in the field of forensic evidence has become more significant in the training for future CI professionals. CI professionals working in the field have often not received this education, nor have they received any structural training or refreshing courses on this topic during their professional careers. The opposite applies for law participants. L students only get basic theoretical education about forensic evidence, but L professionals get the opportunity to follow extensive trainings and refresher courses on forensic evidence. We therefore assume that contrary to our first assumptions, CI students and L professionals who participated in our study have had more theoretical knowledge in the field of forensic evidence than CI professionals and L students. Participants' answers to open questions about how they acquired knowledge on forensic reports also point in this direction. In conclusion, it can be stated that the single fact that someone is a student or professional and may or may not have professional experience in case

Table 6
The mean alleged understanding per evidential strength.

Questions and statements**	Total (M)	Total weak	Total strong
Do you understand the conclusion in the report? (1 I do not understand it at all - 5 I completely understand it)	*4.1**	*3.9*	*4.3*
Do you generally understand the content of the report? (1 I do not understand it at all - 5 I completely understand it)	*4.1***	*3.9*	*4.3*
Do you think there is sufficient information in the report to understand the conclusion? (1 completely insufficient - 5 completely sufficient)	*3.4****	*3.1*	*3.8*

Note: *Significant effect of evidential strength (F(1,1014) = 50.274, p < .001). **Significant effect of evidential strength (F(1,1014) = 42.780, p < .001). *** Significant effect of evidential strength (F(1,1014) = 93.288, p < .001).

Table 5
The mean percentage per answer within each group of participants.

	N	Experience with fingerprint report using CAT (% yes)	Experience with fingerprint report using VLR/NLR (% yes)	Basic understanding of fingerprint conclusions (% yes)	Specific education about fingerprint evidence (% no)
Students	**73**	**23**	**22**	**56**	**34**
legal	31	0	0	51	74
crime investigation	42	52	50	60	19
Professionals	**238**	**74**	**64**	**52**	**18**
legal	116	8	78	65	11
crime investigation	122	85	66	40	25

work is not a good predictor of having knowledge or understanding about a crucial topic in daily practice in the forensic field. In many fields, one becomes an expert by, among others, receiving feedback 'on the job'. This type of structural feedback for professionals is mostly missing in the forensic context and can only be circumstantial since the ground truth is missing. They can be confronted with the opinion of other experts and their interpretation. Further research should study which forms of feedback, such as reviews, are most effective for the forensic field.

In general, differently phrased conclusions with similar evidential strength were assessed differently. The fewest mistakes were made in the assessment of the VLR and NLR conclusions. Most mistakes were made in the assessment of the strong CAT conclusion, using the verbal term 'individualisation'. Although the strength of the weak CAT conclusion was highly underestimated, this conclusion seems to entail information that creates awareness of a level of uncertainty in the evidence. Such information might give rise to a similar awareness when it is used for other conclusion types, particularly strong conclusions. It shows the importance of more explicitly verbalising the uncertainty that is inextricably linked to forensic conclusions. For this to be clear enough, the uncertainty in forensic conclusions needs to be described separately next to the strength of the evidence. As already stated in Van Straalen et al. [4], the uncertainty so clearly recognized in the weak CAT conclusion should be explored to see which aspects of it are useful as the sentence 'it cannot be ruled out that … ' helps to understand this uncertainty.

We believe that regular courses about the interpretation of forensic evidence should be mandatory for law and CI students. Especially for all crime investigation and legal professionals working with forensic evidence, permanent education on the interpretation of forensic reports should be mandatory. Simply being a professional, having had some basic training and having seen and evaluated forensic reports does not make a professional an expert. As also mentioned in Van Straalen et al. [4], forensic advisers with a degree in forensic science should be employed in all criminal justice organisations handling forensic evidence. In addition to providing a neutral explanation about forensic evidence, these advisers can also help signalling whether the knowledge of professionals in their organisations risks being sufficient and up to date. Clear guidelines on the restrictions of this explanation are necessary to keep the chain of evidence transparent and to ensure that the ultimate decision lies with the judge. In addition to improving education about the interpretation of forensic evidence and appointing forensic advisers, the forensic reports themselves can also be improved. As stated before, forensic reports differ greatly between and even within institutions not only in terms of the conclusion type, but also in lay-out and language use [3,34]. Future research should study how specific adjustments to forensic reports can improve their correct interpretation.

This study highlights the importance of receiving correct training and on receiving feedback when it comes to the interpretation of forensic conclusions. Correctly interpreting forensic reports is not automatically learned 'on the job'. Misinterpretation of the evidential value of forensic evidence can have consequences for the correct interpretation of the incriminating or exculpatory nature of the evidence. Incorrect interpretations of evidence can have far-reaching consequences for a fair process of investigation, prosecution and fact finding.

5. Limitations

The topic of this study was the effect of experience and education on the interpretation of forensic conclusions. We used (simplified versions of) fingerprint evidence reports to study this topic. We believe that the outcomes are relevant for diverse types of forensic evidence using a CAT, VLR or NLR conclusion due to the experimental design. Asking face-to-face questions might have given more in-depth insights into the understanding of the reports. Since we only used an online questionnaire, there can never be 100% certainty regarding whether the participants fully understood the questions that were asked.

There was a difference in sample size between the group of student participants and the group of professional participants. This difference did not influence the results. Moreover, the group of law students represented the future professionals of multiple groups of professionals, including public prosecutors, criminal lawyers, and judges. There was a slight change in design after the first study; however, the analyses showed that these changes did not have an effect on the results.

At the start of the study, we did not yet have a complete overview of the exact knowledge and experience of the participant groups. However, the qualitative data from this study on prior knowledge and experience enabled us to check our assumptions.

The participants did not receive additional information about the evidence, suspects, circumstances, or types of crimes. Providing additional information might have given the participants a more realistic experience. However, if we had presented more information, taking part in this study would have taken the participants more time. In addition, given the topic of the study, providing more information might have diverted the focus from the conclusions and we would not have been able to determine precisely what information the participants focused on and what influenced their answers to the questions.

Author contributions

Elmarije van Straalen: Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration.

Christianne 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.

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.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.W. Keijser, de, H. Elffers, Understanding of forensic expert reports by judges, defense lawyers and forensic professionals, Psychol. Crime Law 18 (2) (2012) 191–207, https://doi.org/10.1080/10683161003736744.

[2] J.W. Keijser, H. de, 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.

[3] J.W. Keijser, M. de, 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, https://doi.org/10.1016/j.fsigen.2016.03.006.

[4] E.K. Straalen, C.J. van, Poot, M. de, Malsch, H. Elffers, The interpretation of forensic conclusions by criminal justice professionals: the same evidence interpreted differently, Forensic Sci. Int. 313 (2020), 110331, https://doi.org/10.1016/j.forsciint.2020.110331.

[5] 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, https://doi.org/10.1093/lpr/mgy012.

[6] 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, https://doi.org/10.1016/j.forsciint.2014.04.005.

[7] 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, https://doi.org/10.1371/journal.pone.0213522.

[8] C. Mullen, D. Spence, L. Moxey, A. Jamieson, Perception problems of the verbal scale, Sci. Justice 54 (2) (2014) 154–158, https://doi.org/10.1016/j.scijus.2013.10.004.

[9] K.A. Martire, R.I. Kemp, I. Watkins, M.A. Sayle, B.R. Newell, The expression and interpretation of uncertain forensic science evidence: verbal equivalence, evidence strength, and the weak evidence effect, Law Hum. Behav. 37 (3) (2013) 197–207, https://doi.org/10.1037/lhb0000027.

[10] M. Malsch, M.D. Taverne, H. Elffers, J.W. Keijser, de, P.R. Kranendonk, DNA-Rapporten: Makkelijker Kunnen We Het Niet Maken, Begrijpelijker Wel, Boom Lemma uitgevers, Den Haag, 2013.

[11] 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, https://doi.org/10.1016/j.scijus.2017.02.002.

[12] K.A. Martire, K.N. Ballantyne, A. Bali, G. Edmond, R.I. Kemp, B. Found, Forensic science evidence: Naive estimates of false positive error rates and reliability, Forensic Sci. Int. 302 (2019), 109877, https://doi.org/10.1016/j.forsciint.2019.109877.

[13] 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, https://doi.org/10.1007/s10979-008-9169-1.

[14] A.D. Groot, de, Thought and Choice in Chess, Mouton, The Hague, 1965.

[15] T. Greenhalgh, Intuition and evidence - uneasy bedfellows? Br. J. Gen. Pract. 52 (2002) 395–400.

[16] H.L. Dreyfus, S.E. Dreyfus, Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer, Free Press, New York, 1986.

[17] K.A. Ericsson, The influence of experience and deliberate practice on the development of superior expert performance, in: K.A. Ericsson, P.J. Charness, P. J. Feltovich, R.R. Hoffman (Eds.), The Cambridge Handbook of Expertise and Expert Performance, Cambridge University Press, 2006, pp. 683–703.

[18] M.H. Chi, R. Glaser, M.J. Farr, The Nature of Expertise, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1988.

[19] National Research Council, How People Learn: Brain, Mind, Experience, and School, Expanded Edition, The National Academies Press, Washington, DC, 2000.

[20] A.M. Persky, J.D. Robinson, Moving from novice to expertise and it's implications for instruction, Am. J. Pharmaceut. Educ. 81 (9) (2017) 72–80.

[21] M.T.H. Chi, P.J. Feltovich, R. Glaser, Categorization and representation of physics problems by experts and novices, Cognit. Sci. 5 (2) (1981) 121–152.

[22] N. Dew, S. Read, S.D. Sarasvathy, R. Wiltbank, Effectual versus predictive logics in entrepreneurial decision-making: differences between experts and novices, J. Bus. Ventur. 24 (4) (2009) 287–309, https://doi.org/10.1016/j.jbusvent.2008.02.002.

[23] C.D. Schunn, J.R. Anderson, The generality/specificity of expertise in scientific reasoning, Cognit. Sci. 23 (3) (1999) 337–370.

[24] S. Wiedenbeck, Organization of programming knowledge of novices and experts, J. Am. Soc. Inf. Sci. 37 (5) (1986) 294–299.

[25] S. Lichtenstein, B. Fischhoff, Training for calibration, Organ. Behav. Hum. Perform. 26 (1980) 149–171.

[26] W.K. Balzer, L.B. Hammer, K.E. Sumner, T.R. Birchenough, S.P. Martens, P. H. Raymark, Effects of cognitive feedback components, display format, and elaboration on performance, Organ. Behav. Hum. Decis. Process. 58 (1994) 369–385.

[27] W.K. Balzer, L.M. Sulsky, L.B. Hammer, K.E. Sumner, Task information, cognitive information, or functional validity information: which components of cognitive feedback affect performance? Organ. Behav. Hum. Decis. Process. 53 (1) (1992) 35–54.

[28] W. Remus, M. O'Connor, K. Griggs, Does feedback improve the accuracy of recurrent judgmental forecasts?, in: Paper presented at the Proceedings of the Thirtieth Hawaii International Conference on System Sciences, vol. 3, 1997, https://doi.org/10.1109/HICSS.1997.661557.

[29] C.A.J. Eeden, den van, C.J. Poot, de, P.J. Koppen, van, The forensic confirmation bias: a comparison between experts and novices, J. Forensic Sci. 64 (1) (2019) 120–126, https://doi.org/10.1111/1556-4029.13817.

[30] C. Baber, M. Butler, Expertise in crime scene examination: comparing search strategies of expert and novice crime scene examiners in simulated crime scenes, Hum. Factors 54 (3) (2012) 413–424, https://doi.org/10.1177/0018720812440577.

[31] E.J.A.T. Mattijssen, Forensic Judgments: Validity, Reliability, and Bias, Doctoral Dissertation, Radboud University Nijmegen, 2021, https://hdl.handle.net/2066/233894.

[32] I.E. Dror, Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias, Anal. Chem. (2020), https://doi.org/10.1021/acs.analchem.0c00704.

[33] S.J.W. Willems, C.J. Albers, I. Smeets, Variability in the Interpretation of Dutch Probability 0p1hrases - a Risk for Miscommunication, 2019 arXiv preprint arXiv:1901.09686.

[34] I.E. Dror, G. Hampikian, Subjectivity and bias in forensic DNA mixture interpretation, Sci. Justice 51 (4) (2011) 204–208, https://doi.org/10.1016/j.scijus.2011.08.004.

[35] J. Meeuwissen, R. de Roo, J. Kruithof-van Esch, S. van der Heijden, M. Claushuis, L. van Blijswijk-Kieftenbeld, W. Remijn, Forensic advisers working for all district courts and courts of appeal in the Netherlands: An overview and discussion, J. Forensic Sci. (2023), https://doi.org/10.1111/1556-4029.15385.