Science & Justice 60 (2020) 216–224

Contents lists available at ScienceDirect

Science & Justice

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

Communicating forensic science opinion: An examination of expert reporting practices

Agnes S. Bali[a,*], Gary Edmond[b], Kaye N. Ballantyne[c], Richard I. Kemp[a], Kristy A. Martire[a]

[a] School of Psychology, The University of New South Wales, Sydney, NSW 2052, Australia
[b] School of Law, The University of New South Wales, Sydney, NSW 2052, Australia
[c] Office of the Chief Forensic Scientist, Victoria Police, Melbourne, VIC 3008, Australia

ARTICLE INFO

Keywords:
Forensic science
Expert evidence
Communicating opinion
Uncertainty
Expert reports
Proficiency testing

ABSTRACT

Forensic scientists endeavour to explain complex scientific principles to legal decision-makers with limited scientific training (e.g., police, lawyers, judges, and jurors). Much of the time this communication is limited to written opinions in expert reports. Notwithstanding considerable scientific research and debate about the best way to communicate forensic science opinions, it is unclear how much of the advice has translated into forensic science practice. In conducting this descriptive study, we examined the reporting practices adopted by forensic scientists across a range of forensic science disciplines. Specifically, we used a quantitative content analysis approach to identify the conclusion types and additional information submitted by forensic scientists in proficiency tests during 2016 ("What would be the wording of the Conclusions in your report?"). Our analysis of 500 randomly selected responses in eight disciplines indicated that the conclusion type which has received the most criticism in recent years (categorical statements) remains the preferred means of expression in a clear majority of responses. We also found that the provision of additional information often considered necessary for rational evaluation of the evidence (e.g., information about reliability and validity) was rarely reported. These results suggest limited engagement with recent recommendations and are concerning given the gravity of the legal decisions that hinge on accurate and transparent forensic science communication.

1. Introduction

Forensic scientists must communicate complex scientific principles to legal decision-makers who have little to no scientific training (e.g., police, lawyers, judges, jurors, etc.) [1]. In these situations, clear and accurate communication by forensic scientists is vital, as their scientific findings and opinions can often be pivotal to legal decision-making at numerous stages in criminal justice processes [2]. For example, forensic science findings may be used in investigations (e.g., to identify perpetrators), to inform decisions about charging, prosecution and pleas, to shape legal strategy and questioning at trial, and, of course, to assist with decisions on guilt and non-guilt [3]. Ineffective (i.e., unclear and/or inaccurate) communication at any of these stages has the potential to introduce any number of problems; most conspicuously, threatening the fairness of processes, the accuracy of outcomes, and the efficiency of proceedings [4,5].

Indeed, Garrett and Neufeld [5] provide a salient empirical example of the potential consequences of scientific miscommunication in a legal context. Examining trial transcripts in known wrongful convictions in the U.S., the authors found that 82 (60%) of the 137 transcripts containing expert evidence involved the provision of invalid forensic science testimony. Further, judges routinely admitted invalid testimony without question and prosecutors routinely introduced scientific errors that were not identified or challenged by the defence. While ineffective communication of forensic science evidence is unlikely to have been the sole cause of these wrongful convictions, Garrett and Neufeld's [5] seminal study provides a clear example of how these factors play a role in miscarriages of justice.

It is thus vital for forensic scientists to communicate their scientific opinions in a way that is accurate, clear, and comprehensible to legal decision-makers in criminal justice systems. However, it remains unclear how forensic scientists communicate their scientific opinions to non-scientists in practice, and whether research into improving the accuracy, clarity, and comprehensibility of these opinions is reflected in practice. We aimed to investigate these issues.

* Corresponding author.
E-mail addresses: agnes.bali@unsw.edu.au (A.S. Bali), g.edmond@unsw.edu.au (G. Edmond), kaye.ballantyne@police.vic.gov.au (K.N. Ballantyne), richard.kemp@unsw.edu.au (R.I. Kemp), k.martire@unsw.edu.au (K.A. Martire).

https://doi.org/10.1016/j.scijus.2019.12.005
Received 23 July 2019; Received in revised form 9 December 2019; Accepted 15 December 2019

Table 1
Examples of different conclusion types.

Conclusion type	Example
Categorical	"the crime scene fingerprint matches the suspect's"
Random match probability (RMP)	"one person in 10,000,000 in the population who are not the source would nonetheless match the blood drops"
Likelihood ratio (LR)	"100,000 times more likely if suspect rather than random person is source"
Source probability (SP)	"certain that the suspect is the source"
Likelihood of observed similarity (LOS)	"likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is low"
Strength of support (SOS)	"extremely strong support for the proposition that the samples came from the same source"

Note: examples taken from Thompson and colleagues [21].

1.1. Communication between forensic scientists and legal decision-makers

While it has become increasingly common for legal decision-makers to encounter scientific evidence, opportunities for meaningful dialogue between forensic scientists and decision-makers are generally limited. For example, research has shown that pre-trial consultation between lawyers and forensic scientists is generally underutilised [6,7] and communication between police and forensic scientists is often contingent on factors such as the case type [8]. The provision of expert testimony in court is one circumstance where forensic scientists may have an opportunity to explain their scientific findings in some detail. Although jurors are not able to directly interact with expert witnesses, there are procedures in place in many jurisdictions that enable them to request clarification through the trial judge [9]. Opportunities for additional explanation are usually lost when cases do not proceed to trial, however.

The most common mode of communication between forensic scientists and legal decision-makers is through expert reports and the conclusions expressed therein [10]. The nature and content of these reports is known to be variable [11]. The information included within expert reports may be governed by a number of factors, including but not limited to the nature of the specific case, the purpose of the report, guidelines imposed by the forensic scientist's laboratory, and the forensic scientist's personal experience and preferences [10]. Some reports may provide brief summaries of the forensic scientist's findings, while others may offer comprehensive descriptions of the forensic scientist's entire analytical process [12]. Similarly, some may be succinct interim statements prepared to assist with a police investigation, while others may be formal statements prepared for an anticipated court appearance [10]. Nonetheless, the ultimate goal of these reports is to provide legal decision-makers with accurate and comprehensible, and sometimes timely, information about the scientific findings [2]. Whether forensic scientists are achieving this goal through their written statements is, however, uncertain.

A recent investigation into the general comprehensibility of the opinions within forensic scientists' reports suggests communication through this modality is not as clear as intended. Howes and colleagues [10] investigated the 'readability' of forensic scientists' written conclusions; a construct they defined as "the ease with which something can be read and understood due to the style of writing" and "a prerequisite to comprehensibility" (p. 103). The authors analysed responses submitted by forensic scientists as part of a Collaborative Testing Services Incorporated (CTS) forensic proficiency test in glass analysis. They focused in particular on the section of the test where respondents are asked to express their conclusions as they would in a report, and the readability statistics they computed revealed that at least a tertiary-level education in science was required to understand the written statements with ease. These results suggest that the intended audiences of forensic scientists' reports (legal decision-makers with little to no scientific background) would likely have some difficulty comprehending and accurately applying the information within these reports. But *how* exactly are forensic scientists communicating their opinions?

Howes and colleagues [10] partially investigated this question in a second study, through a qualitative examination of the conclusions in their glass analysis test sample. The authors manually examined features such as the length, text and sentence structure, vocabulary used, and any descriptive information in each conclusion, as a means of evaluating how easily understood they would be by non-scientists. Their findings in this second study corroborated the computer-generated statistics in their first study; the features they hypothesized would increase readability were observed in low proportions across their sample. However, given their exclusive focus on glass analysis, it is unclear whether these findings are generalizable to reporting practices within other disciplines. Further, as the main focus of this study was the readability of the conclusions, Howes and colleagues [10] were not specifically interested in the different types of expressions used to communicate conclusions.

1.2. Different expressions for communicating forensic science conclusions

Forensic scientists may communicate the scientific findings in their reports in a number of different ways (see Table 1 for examples) [13]. They are often tasked with determining whether known and questioned evidence samples (e.g., fingerprints) could have originated from the same source (a suspect) or not, and it has traditionally been accepted for forensic scientists to report their findings categorically [14]. Specifically, a questioned sample is reported as either having originated from the same source as the known sample (an 'identification' or 'match'), a different source as the known sample (an 'elimination' or 'non-match'), or the analysis is determined to be 'inconclusive' [3,14,15]. For certain evidence types (e.g., shoeprints, handwriting, forensic odontology, etc.), these traditional categories have been elaborated to also include the use of statements such as 'could have the same source' and 'probably have the same source' [14,16]. Although there are more categories to choose from, the elaborated categories are still generally used as vague standalone terms rather than as part of an actual informative continuum of probabilities [17].

Alternatively, forensic scientists may adopt quantitative reporting styles incorporating probabilities and statistics to present their findings [14,18]. For example, forensic scientists may choose to report a random match probability (RMP); a numerical estimate of the probability that a 'match' would be obtained between the questioned sample and a sample from a randomly selected member of the reference population (unrelated to the defendant) [3]. Another option is to report a likelihood ratio (LR); a numerical comparison of the likelihood of observing the evidence under two competing hypotheses about the source [19,20]. Forensic scientists may alternatively report a source probability statement (SP); the probability or percent chance that two samples came from the same source [21].

Some forensic scientists may employ procedures generating quantitative results but elect to replace or supplement statistical probabilities with words (i.e., a verbal equivalent; VE). For example, forensic scientists may initially formulate a numerical LR, but then replace or supplement the numerical ratio with a verbal description of strength (ranging from 'weak support' to 'extremely strong support' for one hypothesis over the other [22]). The intention being to facilitate communication. Similarly, a probability or percentage in a source

probability statement may be expressed using verbal instead of numerical terms (e.g., 'moderately probable' or 'practically certain' [21]). Forensic scientists may alternatively report what is essentially the verbal analogue of a random match probability; the likelihood of the observed similarity (LOS) between two samples if they did not have the same source [21]. The verbal labels attached to a LOS statement can range anywhere from 'low' to 'a practical certainty' [21].

1.3. Recent commentary on forensic science communication

The communication practices of forensic scientists have been a subject of increased interest and much scholarly debate in recent years. In 2009, the National Research Council (NRC) questioned communication practices in the forensic sciences. In 2016, President Obama's Council of Advisors on Science and Technology (PCAST) reiterated these concerns about communication as part of an in-depth review of seven feature comparison disciplines. In particular, these authoritative scientific bodies, alongside other scholars [13,18,23–26], have expressed serious concerns about the inadequate communication of uncertainty in reports [27,28].

The opinions expressed by forensic scientists are inherently uncertain [23,25,29]. Forensic scientists can never know, with full certainty, whether the outcomes of their analyses are representative of the ground truth. This is because scientific opinions are necessarily formed using inductive reasoning processes, as opposed to deductive reasoning processes [30]. While deductive reasoning involves using general principles that are known to be true to explain a specific observation, inductive reasoning is limited to using only what can be observed after the fact (evidence samples) to generate general theories about an event (the crime [30]). By definition, any opinion based on inductive inference will always involve some degree of uncertainty [23,30]. It is the degree of this uncertainty that provides legal decision-makers with essential information about the strength or value of the evidence [18]. Indeed, Willis [31] proposed the clear communication of this uncertainty as a part of the forensic scientist's primary responsibility to decision-makers.

Because of emerging expectations around the communication of uncertainty within the forensic sciences, some conclusion types are becoming increasingly controversial. Categorical conclusions are perhaps the most criticised in this respect. That is, absolute categorical statements such as that the samples are 'identical' or 'have the same source' are seen to imply a level of certainty that is unwarranted, scientifically unjustifiable, and indefensible [27,28]. While categorical statements have been defended on the basis that they discourage the undervaluing of scientific evidence [21], the risk (and consequences) of overvaluing the evidence are arguably more problematic. Further, while the elaborated categories used for some evidence types may somewhat address this issue, even these additional categories may be limited in their capacity to convey an accurate level of uncertainty or evidence strength [14]. For example, while the expression '*could have* been the same source' does not eliminate uncertainty entirely, some argue that this expression vaguely suggests an even chance rather than any true indication of the strength of the evidence [16,32]. Similarly, the categorical term 'consistent with' has been shown to lead to ambiguous interpretations of evidence [33].

Source probabilities are also poorly regarded. Source probability statements are seen as problematic because the question of whether a suspect is the source of the evidence is not a scientific one to be decided by a forensic scientist, but a legal one to be decided by a judge and/or jury [21,25,34]. At most, forensic scientists can form opinions about whether two samples originated from the same or different sources. Making statements about the source itself requires the forensic scientist to make unwarranted assumptions about other evidence in the case [21,34]. This practice is widely regarded as well beyond the purview of the forensic scientist's role and expertise [16,21].

The most scientifically robust means of communicating evidence is thought to be the LR [23]. LRs capture the uncertainty inherently associated with scientific analyses through the explicit comparison of two mutually exclusive probabilities: (1) the likelihood of observing the evidence given one proposition (typically the prosecution's proposition that the defendant is the source), and (2) the likelihood of observing the evidence given an alternative proposition (typically the alternate proposition that the defendant is not the source) [22]. By accounting for the relative likelihood of the observations under competing hypotheses, LRs are thought to account for a number of problems that categorical conclusions cannot avoid [14]. For example, LRs can be evaluated on a clear numerical continuum, while the boundaries within a categorical reporting system (i.e., what decides whether a conclusion is a 'match', 'non-match', or 'inconclusive') are generally arbitrary and poorly explained [3,30]. On the other hand, there are known issues with generating numerically accurate LRs [21,35,36] as well as concerns about how well they are understood by decision-makers [37–39].

Alongside discussion of the utility of varying conclusion types, there has also been increased interest in the additional content (or lack thereof) within forensic scientists' reports. For example, in light of recent research revealing the limited scientific foundations for a number of forensic sciences disciplines, forensic scientists have been encouraged to supplement their opinions with the provision of information about the reliability and validity of the analytical methods they employ [28]. Scholars have also recommended including information about other types of potential limitations to the analysis and the final conclusion (e.g., limitations to the specific forensic scientist's proficiency) [40,41]. Further, the tendency for forensic scientists to pose their conclusions as though they were fact rather than simply an opinion has been admonished, and transparency surrounding uncertainty has been encouraged [27,28,41]. Ultimately, forensic scientists have been urged to give much greater consideration to communicating the limits of their practice, and some codes of conduct (e.g., the Australian Federal Harmonised Expert Witness Code of Conduct) and practice directions (e.g., the Supreme Court of Victoria Practice Note on Expert Evidence in Criminal Trials) require it [42].

While there has been no consensus on a precise set of communication standards, attentive scholars have identified a range of practices warranting revision. Many scholars and all of the major scientific reviews have questioned some types of expression and recommended the provision of additional information [14,27,28,41]. Yet there is little information about how much (if at all) this informed advice is translating into reporting practices and testimony. The present descriptive study aimed to address this gap by auditing the current communication practices in a range of forensic science disciplines. Specifically, we aimed to capture the precise conclusion types and content of reports produced by forensic scientists. Following past research [10,43,44], we examined this question by surveying responses submitted by forensic scientists in CTS proficiency tests. While previous research using this data source has focused on the accuracy (error rates) [43,44] or readability [10] of the test responses, we were specifically interested in what information was communicated within the responses (and how), irrespective of these factors. An evaluation of such responses provides substantial insight into the reporting practices used by forensic scientists.

2. Method

2.1. Sample

Our sample of forensic conclusions was sourced from forensic proficiency test results uploaded online by the international test provider *Collaborative Testing Services Incorporated* (CTS). Proficiency tests are commonly used by laboratories to assess the performance of their forensic scientists [43]. These tests typically involve the completion of tasks that are intended to simulate real casework but for which the ground truth is known. Regular participation in forensic proficiency

Table 2
Coding categories and examples.

Coding Category	Example
Categorical conclusion type	(Traditional) "the questioned sample matches the known sample" (Elaborated) "the questioned sample probably has the same source as the known sample"
Random match probability (RMP) conclusion type	"the chance of observing the evidence if it originated from an unrelated individual is 1 in 20"
Likelihood ratio (LR) conclusion type	(Numerical) "the observation of the evidence is 1000 times more likely if the sample came from the defendant than if it came from an another source" (Verbal) "there is very strong support for the hypothesis that the evidence came from the suspect rather than if it did not come from the suspect"
Source probability (SP) conclusion type	(Numerical) "there is a 90% chance that the suspect is the source" (Verbal) "it is highly probable that the suspect is the source"
Likelihood of observed similarity (LOS) conclusion type	"the likelihood of observing this amount of corresponding detail when the evidence did not come from the same source is low"
Strength of support (SOS) conclusion type	"extremely strong support for the proposition that the samples came from the same source"
Alternative explanations for the evidence	"this result might also have been observed if…"
Reliability (of the analysis methods used, of the forensic scientist, etc.)	"this procedure is known to have an error rate of…"
Validity (of the analysis methods used, of the discipline, etc.)	"this analysis method has been empirically validated"
Potential limitations to the analysis and final opinion	(General) "the method used does not account for…" (Specific) "these results may be affected by the fact that the sample was particularly deteriorated"
Explicit opinion statement	"it is the opinion of this examiner that…"
Methods	(Vague) "using standard laboratory procedures" (Specific) "after visual inspection and manual comparison"
Reasoning or justification for the final opinion	"the questioned sample corresponds to the known sample in design, physical size, wear and two randomly acquired characteristics. Therefore…"
Explanation of jargon or scientific terminology	"a conclusion of 'is eliminated' indicates that…"

testing is a requirement for accreditation to ISO/IEC 17025, an international standard for laboratory competency [44]. Proficiency test results are therefore potentially a rich source of information about what is considered best practice within the forensic sciences [10,43,44].

CTS regularly publishes summary reports containing the results of their tests on their website (https://cts-forensics.com/index-forensics-testing.php) as well as all individual responses to that test. CTS tests are available for purchase by institutions, practitioners, and the general public, at costs ranging from approximately US$160 to US$845 (depending on the discipline) per administration. The data for all respondents to a given test are included in the publicly available online repository.

When this study was undertaken, the 2016 summary reports were the most recent available across all of the disciplines. The study sample thus comprised 500 randomly selected conclusions from the 2016 summary reports for eight different forensic science disciplines: fibres analysis, firearms examination, glass analysis, handwriting examination, paint analysis, questioned documents examination, shoeprint impression evidence, and tool marks examination. Decisions regarding which disciplines to include were firstly determined by whether the test for a discipline included the specific question: "What would be the wording of the Conclusions in your report?". As we chose to focus on feature comparison procedures, we then excluded disciplines that did not include a relevant example.

2.2. Procedure

All conclusions from each of the summary reports submitted to CTS during 2016 in the disciplines specified above (two summary reports each in firearms examination, paint analysis, and toolmarks examination, one summary report each in fibres analysis, glass analysis, handwriting examination, questioned documents examination, and shoeprint impression evidence) were copied into an Excel spreadsheet and assigned a unique ID number. The random number generator function in Excel was then used to generate 500 random numbers. The 500 conclusions with the corresponding ID numbers were then copied into a separate spreadsheet for audit: 43 conclusions in fibres analysis, 121 in firearms examination; 33 in glass analysis; 52 in handwriting examination; 39 in paint analysis; 59 in questioned documents examination; 64 in shoeprint impression evidence; and 89 in toolmarks examination.

A quantitative content analysis framework was used to analyse the conclusions in the final sample. Content analysis has been defined as "the systematic, objective, quantitative analysis of message characteristics" (p. 1) and involves coding verbal or written communications for various characteristics [45]. In this study, coding categories for different conclusion types and content were derived from: (1) a review of the literature; and (2) a preliminary examination of 50 conclusions to determine if any coding categories had not been accounted for. These 50 conclusions were separate from the final sample of 500.

The final coding categories consisted of six conclusion types: categorical conclusions, LRs, RMPs, likelihood of observed similarity statements, strength of support statements, and source probabilities. These categories were based largely on the work of Thompson and colleagues [14,21,34]. In addition, we also coded eight types of content: description of the analysis methods used, information about method reliability, information about method validity, limitations to the methods or conclusions, explication that the conclusion is an opinion, reasoning or justification for the final conclusion (e.g., information about feature frequency or similarities), any alternative explanations for the results, and additional explanation of jargon or scientific terminology; as per recommendations made by PCAST [28], NRC [27], AFSP [22], and the like. Most categories were binary coded ('absent' or 'present'), however some of the categories incorporated additional coding levels. Categorical conclusions, for example, were coded as either 'absent', 'only traditional type present', 'only elaborated type present', or 'both traditional and elaborated types present'. Similarly, the inclusion of analysis methodology was coded as either 'absent', 'vague description present', or 'specific description present'. See Table 2 for a summary of the coding categories and examples. A coding sheet with full descriptions of each category, their coding levels and coding rules was generated before coding the final sample commenced.

A member of the research team then used the coding sheet to code responses under the 'Conclusions' section of the CTS Summary Reports. As mentioned, this section contained responses to the question: "What would be the wording of the Conclusions in your report?" We noted that respondents occasionally included extra information about their opinions within the 'Additional Comments' section of the Summary Reports. This section of the test is framed as a space for any supplementary comments or feedback the respondents wish to give to CTS as a

Table 3
Proportions (%) of conclusion types present by discipline.

	Fibres analysis	Firearms examination	Glass analysis	Handwriting examination	Paint analysis	Questioned documents	Shoeprint impression	Toolmarks examination	Total
Categorical									
Traditional	39.5	95.9	18.2	55.8	35.9	93.5	66.7	100.0	70.6
Elaborated	32.6	–	36.4	13.5	41.0	–	–	–	12.4
Both	27.9	1.7	42.4	17.3	7.7	–	12.5	–	10.9
Random match probability	–	–	–	–	–	–	–	–	–
Likelihood ratio									
Numerical	–	–	–	–	–	–	–	–	–
Verbal	–	–	6.1	–	–	–	–	–	0.5
Both	–	–	–	–	–	–	–	–	–
Likelihood of observed similarity	–	0.8	3.0	3.8	–	–	–	–	0.9
Source probability									
Numerical	–	–	–	–	–	–	–	–	–
Verbal	–	–	–	–	–	–	–	–	–
Both	–	–	–	–	–	–	–	–	–
Strength of support	2.3	0.8	12.1	11.5	7.7	6.5	4.2	–	4.6

Note. '–' = 0% proportion observed.

test provider, and there was considerable variation in the nature of these comments. For example, comments ranged from complaints about the test samples to full reproductions of laboratory testing and reporting policies. Because of this variability and the understanding that these could be considered peripheral details that would not be included within the wording of the conclusions in the respondents' reports, we did not include this section in our analysis.

Following completion of the coding process, we used a 'split-half' technique to check the likelihood that the sample size was sufficient to be representative of the full sample of conclusions in the CTS summary reports [46]. This involved splitting the 500 conclusions multiple times into sub-samples of approximately half, and comparing the results of each half against the other.

Following conventional practice [47], a randomly selected subset of our sample (100 conclusions or 20% of the sample) was also coded by two independent raters (doctoral candidates) who were blind to the likely study outcomes. These raters were initially trained by coding ten practice conclusions using the same coding sheet that had been used by the primary coder. They then used the coding sheet to independently code the same 100 conclusions. Inter-rater agreement was then computed by comparing the degree of agreement between all three raters on each category.

3. Results

3.1. Split-half analysis

The 'split-half' technique was used to examine whether our sample size was sufficient [46]. If comparison of the split dataset indicates a large degree of difference between the conclusions that can be drawn from the two halves, it would suggest the sample size is not sufficient to draw general conclusions about the population of interest [46]. Comparisons of our split dataset indicated no significant differences between the sub-samples for each of the five times this technique was performed. This indicates that the sample size is sufficiently representative.

3.2. Inter-rater agreement

We computed Krippendorff's α coefficients [46] for each of the conclusion type and additional content coding categories to assess inter-rater agreement. Krippendorff's α is a commonly used measure of reliability for content analysis and is accommodating of any number of raters, categories, and observations [48]. Values of $\alpha = 0.80$ or higher are considered a strong level of agreement, while any value lower than $\alpha = 0.667$ is considered to be inadequate agreement [45,46,49].

3.2.1. Conclusion types

Inter-rater agreement was good for categorical conclusions (Krippendorff's $\alpha = 0.80$), source probability statements ($\alpha = 0.85$), and strength of support statements ($\alpha = 0.81$). Inter-rater agreement was excellent for likelihood of observed similarity ($\alpha = 1.00$), RMP ($\alpha = 1.00$), and LR ($\alpha = 1.00$) statements.

3.2.2. Additional content

Inter-rater agreement was good for the provision of reasoning or justification for final conclusions ($\alpha = 0.80$), method reliability ($\alpha = 0.85$), and method validity ($\alpha = 0.86$), and adequate for inclusion of alternative explanations for the results ($\alpha = 0.78$), description of analysis methods ($\alpha = 0.77$), limitations to methods or conclusions ($\alpha = 0.76$), and additional explanation of jargon or scientific terminology ($\alpha = 0.71$). Inter-rater agreement was excellent for explicit opinion statements ($\alpha = 1.00$).

3.3. Conclusion types used by respondents

Overall, traditional categorical statements (e.g., 'match', 'identification', 'elimination', etc.) was by far the most prevalent conclusion type observed, and was coded as present in 70.6% of conclusions. Elaborated categorical statements (e.g., 'could have originated from the same source') were coded as present in 12.4% of conclusions, and 10.9% of conclusions included both traditional and elaborated categorical statements. Verbal LRs (e.g., 'it is 100 times more likely that the samples have the same source rather than a different source') and likelihood of observed similarity statements (e.g., 'correspondence between the samples is low') were observed in less than 1% of conclusions (0.5% and 0.9%, respectively). Strength of support statements (e.g., 'there is moderately strong support for the proposition…') were coded as present in 4.6% of conclusions. RMPs and source probabilities were not observed in any of the conclusions included in the analysis. See Table 3 for a breakdown of the proportions of conclusion types by discipline.

3.4. Additional content provided by respondents

The provision of some reasoning or justification for the final conclusion (e.g., 'footwear impression corresponds to the left shoe in outsole design, physical size, wear and two randomly acquired

Table 4
Proportions (%) of additional content types present by discipline.

	Fibres analysis	Firearms examination	Glass analysis	Handwriting examination	Paint analysis	Questioned documents	Shoeprint impression	Toolmarks examination	Total
Alternatives	18.6	13.2	21.2	–	30.8	6.5	–	–	11.6
Reliability	–	–	–	–	2.6	–	–	–	0.3
Validity	–	–	–	–	2.6	–	–	–	0.3
Limitations									
General	7.0	4.1	6.1	3.8	5.1	3.2	–	–	3.8
Specific	–	0.8	–	13.5	–	3.2	–	–	2.3
Methods									
Vague	7.0	14.9	–	3.8	10.3	19.4	12.5	9.6	10.4
Specific	18.6	23.1	24.2	3.8	15.4	19.4	4.2	28.85	18.7
Opinion statement	4.7	6.6	12.1	17.3	10.3	19.4	8.3	5.8	9.6
Reasoning	67.4	43.8	60.6	30.8	46.2	54.8	41.7	40.4	46.6
Term explanation	2.3	0.8	–	1.9	–	–	1.6	–	0.8

Note. '–' = 0% proportion observed.

characteristics, and therefore…') was the most prevalent content type, coded as present in 46.6% of conclusions. Specific information about analysis procedures (e.g., 'using visual inspection, manual comparison, and elemental analysis') was observed in 18.7% of conclusions, and vague information (e.g., 'using standard laboratory procedures') was observed in 10.4% of conclusions. Possible alternative explanations (e.g. 'this result might also be observed if…') were coded as present in 11.6% of conclusions, and explicit opinion statements (e.g., 'it is my opinion that…') were observed in 9.6% of conclusions. The provision of information related to the reliability (e.g., 'this method has an error rate of…') or validity (e.g., 'this method has been empirically validated') of the analysis was observed in less than 1% of conclusions, and general and/or specific limitations (e.g., 'the sample was deteriorated') were observed in less than 5% of conclusions (3.8% and 2.3%, respectively). Additional explanation of scientific terminology or jargon (e.g., 'a conclusion of "is eliminated" indicates that…') was coded as present in less than 1% percent (0.8%) of conclusions. See Table 4 for a summary of the proportions of additional content types by discipline.

4. Discussion

In conducting this descriptive study, we intended to audit the reporting practices adopted by forensic scientists in reports intended for decision-makers with limited scientific knowledge or training. Although there has been much commentary and debate about the most scientifically accurate means of expressing forensic science conclusions, it is still unclear how much of this progress has translated into practice.

Our examination of the conclusion types present in the 2016 forensic proficiency test results indicates that, in fact, few of the scientifically-informed recommendations appear to have permeated forensic science practice. We found that the conclusion type which has received the most criticism in recent years (categorical statements) was the preferred expression in the overwhelming majority of responses. On the other hand, the conclusion type that is considered the most scientifically robust (the LR) was used in only two responses. RMPs and source probabilities were not observed in any of the responses, while strength of support and likelihood of observed similarity statements were observed in less than 5% of all responses.

Our analysis indicated similar deficiencies with respect to the additional content within the responses. Information about methodological reliability and validity, limitations (specific or otherwise), and further explanation of scientific jargon was observed in less than 10% of all responses. Just under 12% of responses offered some consideration of alternative explanations for the findings and included the explicit statement that the conclusion was the opinion of the forensic scientist. Still, a substantial number of forensic scientists' responses (almost 50%) included some reference to the reasoning behind their opinions and some description of the analysis methods used (29% prevalence). These findings suggest some – though limited – engagement with recent commentary and recommendations.

These results provide novel insight into the conclusion types and additional content being reported in proficiency tests, and presumably in practice, across a number of different disciplines. However, we are able to compare some of our findings to those of Howes and colleagues [10]. Some of the coding categories used by Howes and colleagues [10] in their examination of glass analysis proficiency test data do appear to somewhat overlap with ours. In particular, the coding categories they labelled relevant to 'findings from comparison' and 'inference from findings' appear to correspond with what we have labelled as categorical statements (e.g., "was consistent with", "could have originated from the same source", etc.; p. 106). Consistent with what we observed in our glass analysis sample (and across the disciplines in our sample), Howes and colleagues [10] coded these statements as present in high proportions within their sample of glass analysis conclusions (ranging from 72% to 81% prevalence).

We also observed similarities with Howes and colleagues [10] with respect to the additional content within the glass analysis responses. We observed similarly low proportions of explicit opinion statements (12% in our glass analysis sample; 5% and 10% in Howes and colleagues' sample [10]) and limitations (6% in our sample; 8% and 10% in their sample), and similarly high proportions of some type of reasoning being present (61% in our sample; 49% and 66% in their sample). However, Howes and colleagues [10] observed almost double the amount of considering alternatives (38%) to what we observed in our glass analysis sample (21%). It is unclear why this was the case, however it may be due to differences in the way this category was defined and coded across the studies (e.g., differences in what level of detail qualified as the presence of an alternative explanation, etc.).

The general ubiquity of categorical statements across the disciplines (and across studies; [10]) is of concern given that the use of terms such as 'match', 'identification', or 'individualisation' are considered to encourage legal decision-makers to overvalue the impressions of forensic scientists [34,50]. There are a number of reasons that may explain why categorical conclusions are still pervasive in forensic science reporting. First, forensic scientists may simply not be aware of the commentary and controversy surrounding the use of this conclusion type. However, given this issue became prominent almost ten years ago [27] and has only gained momentum since [28,51], any lack of familiarity or response is concerning and raises serious questions about standards of professionalism and training [41]. The second, perhaps more likely, explanation is that forensic scientists may be constrained by reporting policies that are put in place by their laboratories [10]. If forensic scientists are required to use outdated forms of expression by their institutions, this suggests that reform of forensic science communication practices is required at the highest levels [28].

There are also difficulties associated with adopting the reporting

approaches that are more highly regarded. As discussed, the LR is considered the most logically and scientifically appropriate means for forensic scientists to communicate their conclusions. One advantage of the LR is that it is able to numerically communicate the relative probability of observing similarities and differences between samples under two different hypotheses [21,23]. However, it is no easy task to accurately generate these numerical ratios. That is, there are very few disciplines (e.g., DNA analysis) in which LRs can be calculated using aleatory databases and statistical modelling techniques [17,21]. In many cases values are unavoidably generated from the non-systematic experience and recollection of practitioners [52]. In these cases the estimates may be 'numbers from nowhere' [36] and could misleadingly create the illusion of a mathematically precise quantification of evidential weight when in reality they have no rigorous empirical foundation. Indeed, evidence obtained from court-going forensic document examiners suggests that forensic scientists can learn the frequency of occurrence for handwriting features more accurately than novices, but that these estimates deviate from the likely 'true' values by approximately 20 percentage points [53]. There is also some ambiguity surrounding exactly what information can and should be included in the formulation of the numerical ratio [35], and whether certain values (e.g., lab error rates, etc.) may be better represented on their own rather than incorporated into the final ratio [25]. Thus the absence of a solid empirical basis for generating values in many disciplines and questions regarding the validity of self-generated estimates may reasonably account for delays in adopting LR-based reporting methods.

Setting aside the issues specific to the use of LRs, there are also general concerns associated with the use of probabilistic and statistical expression types (LRs, RMPs, etc.) to communicate information to laypeople. Specifically, that probabilistic expressions undermine legal decision-makers' comprehension of the evidence [3,50,54]. Indeed, empirical studies in both legal and non-legal contexts suggest decision-makers find it is a challenge to comprehend probabilities and statistics [14,54–57]. Even forensic scientists themselves (along with judges and lawyers) have been shown to make numerous mistakes in their interpretations of LRs [37]. Some have suggested that the use of VEs may resolve this issue by using 'universal' language. However, research has found that forensic scientists and legal decision-makers seem to have disparate interpretations of designated VE terms, and VEs may therefore not adequately convey the intended levels of strength either [38,39,58,59]. As such, moving toward recommended reporting practices may not actually be as conducive to the improved understanding and use of the evidence as hoped. It remains unclear how the need for accuracy in communication can be reconciled with lay comprehension.

Perhaps the problem is not so much that forensic scientists are using categorical statements, but that the statements do not seem to be being qualified using relevant and informative additional content that can give shape to the associated uncertainty [24,41]. Mitchell and Garrett [60] recently showed that providing lay participants with information about the number of errors the expert made in their proficiency testing significantly influenced the weight participants assigned to fingerprint expert testimony. These results suggest that additional information about alternative explanations and/or limitations to the conclusion can have a powerful impact on how the evidence is perceived by decision-makers. In light of this result, it is somewhat concerning that the provision of information about a possible alternative explanation was observed in under 12% of responses in our data and that the inclusion of potential limitations was observed in fewer than 10% of conclusion statements.

It is of particular note that forensic scientists in our analysis appear to have almost entirely omitted additional information relevant to the reliability and validity of the analysis methods used. It is well accepted within the scientific community that the value of a forensic scientist's findings are constrained by the accuracy and capacity of their methods [28]. Indeed, PCAST [28] has recommended that forensic scientists be transparent about the empirical support for their procedures (even if to state that there is none) to avoid implying a higher degree of certainty than is warranted given the analytical methods available. That this information may not be provided by forensic scientists is particularly concerning given that two disciplines included in this study (firearms examination and shoeprint/footwear impressions) were recently found to have questionable scientific foundations [28]. Although we note that the conclusions reached by PCAST have not been universally accepted (for a summary and critical analysis see [61]). It is possible that individual forensic scientists did not provide pertinent information as part of their CTS response because some institutions use standardised reporting templates that already incorporate it. Thus, the status of validity and reliability information in forensic science reports remains somewhat unclear.

It is reassuring that a substantial number (almost 50%) of conclusions in our analysis involved the provision of some reasoning or justification for the conclusion. Psychological research has consistently shown that it is often not enough to assert that something is the way it is; people are better able to remember, understand, and appropriately use information when they are given an explanation for *why* it is the way it is [50]. This suggests that legal decision-makers are likely to have a better understanding of the value of the link between two samples when they are given an idea of the similarities and discrepancies between the two. It could be argued, however, that these justifications mean little in the absence of information about the reliability and validity (and their limits) of the analysis the forensic scientist has performed in order to reach their opinion. Decision-makers may thus subjectively feel they have a better understanding of the evidence, but this does not necessarily mean that they will be able to evaluate it in a sensible manner without a better understanding of its limits.

Overall, our results show that forensic scientists may not be providing enough information for legal decision-makers to be able to appropriately use their findings. While there may be some systemic factors which influence the information a forensic scientist will include in their proficiency test responses, ultimately, the respondents in our analysis were asked to specify the wording of the conclusion they would submit in a regular report. Assuming that the responses provided in these proficiency tests are actually what forensic scientists would include in their reports (which is a reasonable expectation given the purpose of these tests), then as of 2016 these responses were rarely of the form currently recommended (LRs). Moreover, they rarely, if ever, included all the caveats and additional information necessary for the transparent, modest, and accurate communication of results [41].

5. Limitations

There are some important limitations to this analysis. First, some responses that were included in our examination may not have been submitted by trained forensic scientists. That is, CTS summary reports include the responses of all individuals who complete the test(s) in question and CTS does not restrict their services to only forensic scientists from accredited or accreditation-seeking laboratories. The tests are made available to anyone who is willing and able to afford the test. CTS does not provide any specific demographic information for their test respondents, so there is also no way to identify and exclude responses that were not submitted by trained forensic scientists. Even so, it is reasonable to assume that the number of non-forensic scientists who complete these tests is small as, although the financial cost varies depending on the type of test, it is prohibitive for most (if not all) disciplines.

Second, some forensic scientists might have omitted information from their CTS test responses that would normally be included in standardised templates used by their institutions. The reporting policies of some institutions require forensic scientists to simply insert an opinion statement into a pre-existing template which describes standardised laboratory procedures, known error rates, and other general

information relevant to that particular forensic science discipline. There may also be little incentive to include any information considered non-essential, given that these conclusions are generally not the focus of CTS' grading system. Further, depending on the discipline, CTS may also place constraints on the response that can be given. As a result, some of the conclusions in our analysis may not represent the full extent of information that the decision-maker might receive in a report. Not all institutions employ the use of standardised templates, however we have no way of determining how many respondents in our sample might have responded based on such practices. Even so, it would be reasonable to assume that most forensic scientists would not expect CTS to have knowledge of their institution-specific templates and might adapt their responses to be more reflective of what they believed would be received well by the testing service. Further, a number of the additional content categories we chose to examine are case-specific (e.g., specific limitations to the analysis) and would therefore remain unaccounted for by the general information within standardised templates. The results of our analysis showed that even these types of details were rarely addressed by forensic scientists. The use of templates is unlikely to wholly account for the general sparseness of the conclusions in the current study.

Third, our entire dataset was comprised of forensic proficiency test results, which have a reputation for being too easy and a poor reflection of real casework [44]. These factors are perhaps not so relevant to us, however, given that we were not interested in whether forensic scientists reached the correct conclusion, but simply how they chose to communicate whatever conclusion they reached. There are also concerns that because proficiency tests are not administered 'blind', forensic scientists may respond differently to how they would in real casework. However, we would argue that even if respondents were expressing their conclusions differently to how they normally might, it is reasonable to anticipate that they would be more rather than less diligent in the completion of CTS testing on the basis that it is relevant to individual and laboratory level certification and accreditation.

Finally, we were only able to analyse certain forensic science disciplines because of our inclusion criteria. That is, we were only able to include the results of tests which included the specific question asking respondents to express their conclusions as they would in their report. We may have observed completely different proportions of particular coding categories if other feature-comparison disciplines had been able to be included in the analysis. For example, we may have observed a much higher proportion of LRs if DNA analysis had been one of the disciplines included in our sample. This is because the LR has long been the common approach for reporting the results of comparisons involving mixed DNA samples [34]. Similarly, we might have observed a greater proportion of statistical statements in general if a discipline such as forensic voice comparison was able to be included in this analysis, as this is another example of a discipline which has established databases and statistical techniques that can be used to accurately calculate numerical values and ratios [21,62]. Still, the disciplines included within our analysis represent a substantial and varied portion of forensic science disciplines. It would nonetheless be valuable for future research to examine the reporting practices across a broader range of disciplines.

6. Conclusion

Although there has long been scholarly commentary and debate regarding the state of reporting within the forensic sciences, until now the extent to which this debate was actually affecting practice was unclear. This descriptive study provides valuable insights into the status of expert communication in the forensic sciences. Through a study of proficiency tests, we found that questionable communication practices remain prevalent in the conclusions used by forensic scientists. In considering the prevalence of non-scientific reporting practices it is important to consider whether forensic scientists are being afforded the means to move away from these styles of reporting (i.e., empirical bases for statistical conclusions and institutional support). Further, there may be a general reluctance by forensic scientists to shift to alternative reporting practices that may be more scientifically robust but problematic for comprehension by non-scientists. Irrespective, our analysis showed that forensic scientists volunteer little information to qualify their absolute conclusions. This is likely to leave their intended audiences - legal decision-makers with limited scientific training - in a position where they may significantly overvalue the evidence given the known but undeclared limits associated with all inferential scientific techniques. Such practices are concerning given the gravity of the legal decisions that hinge on accurate and transparent forensic science communication.

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.

Acknowledgements

The authors would like to thank the two independent raters (SS and JLG) for their assistance in carrying out this research.

Role of the funding sources

This work was supported by an Australian Government Research Training Program (RTP) Scholarship and an Australian Research Council (ARC) Linkage Project (LP160100008).

References

[1] L.M. Howes, The communication of forensic science in the criminal justice system: a review of theory and proposed directions for research, Sci. Justice 55 (2) (2015) 145–154, https://doi.org/10.1016/j.scijus.2014.11.002.
[2] T. Rothwell, Presentation of expert forensic evidence, in: P. White (Ed.), *Crime Scene to Court: The Essentials of Forensic Science*, The Royal Society of Chemistry, Cambridge, UK, 2010, pp. 507–532.
[3] L.M. Howes, N. Kemp, Discord in the communication of forensic science: can the science of language help foster shared understanding? J. Lang. Soc. Psychol. 36 (2017) 96–111, https://doi.org/10.1177/0261927x16663589.
[4] S. Cordner, R.v. Klamo, An example of miscommunication and misunderstanding of expert evidence where the conviction was overturned, Aust. J. Forensic Sci. 44 (4) (2012) 323–331, https://doi.org/10.1080/00450618.2012.691551.
[5] B.L. Garrett, P.J. Neufeld, Invalid forensic science testimony and wrongful convictions, Virginia Law Review (2009) 1–97.
[6] K. Cashman, T. Henning, Lawyers and DNA: issues in understanding and challenging the evidence, Curr. Issues Criminal Justice 24 (1) (2012) 69–83, https://doi.org/10.1080/10345329.2012.12035945.
[7] R. Wheate, Australian forensic scientists: a view from the witness box, Aust. J. Forensic Sci. 40 (2) (2008) 123–146.
[8] S.F. Kelty, R. Julian, A. Ross, Dismantling the justice silos: avoiding the pitfalls and reaping the benefits of information-sharing between forensic science, medicine and law, Forensic Sci. Int. 230 (2013) 8–15, https://doi.org/10.1016/j.forsciint.2012.10.032.
[9] B.M. Dann, V.P. Hans, D.H. Kaye, *Testing the Effects of Selected Jury Trial Innovations on Juror Comprehension of Contested mtDNA Evidence: Final Technical Report*, National Institute of Justice, Office of Justice Programs, US Department of Justice, 2004.
[10] L.M. Howes, K.P. Kirkbride, S.F. Kelty, R. Julian, N. Kemp, Forensic scientists' conclusions: how readable are they for non-scientist report-users? Forensic Sci. Int. 231 (1–3) (2013) 102–112, https://doi.org/10.1016/j.forsciint.2013. 04.026.
[11] B. Found, G. Edmond, Reporting on the comparison and interpretation of pattern evidence: recommendations for forensic specialists, Aust. J. Forensic Sci. 44 (2012) 193–196, https://doi.org/10.1080/00450618.2011.644260.
[12] G. Edmond, S. Carr, E. Piasecki, Science friction: streamlined forensic reporting, reliability and justice, Oxford J. Legal Stud. 38 (4) (2018) 764–792.
[13] G. Jackson, Understanding forensic science opinions, Handbook Forensic Sci. (2009) 419–455, https://doi.org/10.4324/9781843927327.ch16.
[14] W.C. Thompson, E.J. Newman, Lay understanding of forensic statistics: evaluation of random match probabilities, likelihood ratios, and verbal equivalents, Law Hum Behav. 39 (4) (2015) 332–349, https://doi.org/10.1037/lhb0000134.
[15] S.A. Cole, Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States, Law Probab. Risk 13

(2014) 117–150.
[16] W.C. Thompson, S.A. Cole, Psychological aspects of forensic identification evidence, in: M. Costanzo, D. Krauss, K. Pezdek (Eds.), *Expert Psychological Testimony for the Courts*, Lawrence Erlbaum Associates, Mahwah, NJ, 2007, pp. 31–68.
[17] B. Robertson, G.A. Vignaux, C.E. Berger, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, John Wiley & Sons, West Sussex, UK, 2016.
[18] J. Koehler, M. Saks, Individualization claims in forensic science: still unwarranted, Brooklyn Law Rev. 75 (2010) 1187.
[19] J.J. Koehler, On conveying the probative value of DNA evidence: frequencies, likelihood ratios, and error rates, Univ. Colorado Law Rev. 67 (1996) 859–886.
[20] J.J. Koehler, When are people persuaded by DNA match statistics? Law Hum Behav. 25 (5) (2001) 493–513, https://doi.org/10.1023/a:1012892815916.
[21] W.C. Thompson, R.H. 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.
[22] Association of Forensic Science Providers [AFSP], Standards for the formulation of evaluative forensic science expert opinion, *Sci. Justice* 49 (3) (2009) 161–164, https://doi.org/10.1016/j.scijus.2009.07.004.
[23] C. Aitken, P. Roberts, G. Jackson, Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Royal Statistical Society, London, 2010.
[24] G. Edmond, M.B. Thompson, J.M. Tangen, A guide to interpreting forensic testimony: scientific approaches to fingerprint evidence, Law Probab. Risk 13 (1) (2014) 1–25.
[25] G. Jackson, D.H. Kaye, C. Neumann, A. Ranadive, V.F. Reyna, *Communicating the Results of Forensic Science Examinations: Final Technical Report*, National Institute of Standards and Technology, United States, 2015.
[26] M. Redmayne, P. Roberts, C. Aitken, G. Jackson, Forensic science evidence in question, Criminal Law Rev. 5 (2011) 347–356.
[27] National Research Council [NRC], *Strengthening Forensic Science in the United States: A Path Forward*, The National Academies Press, Washington, D.C., 2009.
[28] President's Council of Advisors on Science and Technology [PCAST], *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods*, Executive Office of The President's Council of Advisors on Science and Technology, Washington D.C., 2016.
[29] C.E. Berger, J. Buckleton, C. Champod, I.W. Evett, G. Jackson, Expressing evaluative opinions: a position statement, Sci. Justice 51 (1) (2011) 1–2, https://doi.org/10.1016/j.scijus.2011.01.002.
[30] C.E. Berger, Criminalistics is reasoning backwards, *Nederlands Juistenblad* 85 (2010) 784–789.
[31] S. Willis, Forensic science, ethics and criminal justice, in: J. Fraser, R. Williams (Eds.), *Handbook of Forensic Science*, Willan, Devon, UK, 2009, pp. 523–545.
[32] C. Aitken, An introduction to a debate, Law Probab. Risk 11 (4) (2012) 255–258, https://doi.org/10.1093/lpr/mgs014.
[33] R. Ross, K. Kramer, K.A. Martire, Consistent with: what doctors say and jurors hear, Aust. J. Forensic Sci. 1–8 (2017), https://doi.org/10.1080/00450618.2017.1324583.
[34] W.C. Thompson, How should forensic scientists present source conclusions? Seton Hall Law Rev. 48 (2018) 773.
[35] K.A. Martire, G. Edmond, D.J. Navarro, B.R. Newell, On the likelihood of "encapsulating all uncertainty", Sci. Justice 57 (2017) 76–79, https://doi.org/10.1016/j.scijus.2016.10.004.
[36] D.M. Risinger, Reservations about likelihood ratios (and some other aspects of forensic 'Bayesianism'), Law Probab. Risk 12 (1) (2013) 63–73, https://doi.org/10.2139/ssrn.2020052.
[37] J. de Keijser, 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.
[38] 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.
[39] 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.
[40] G. Edmond, Forensic science evidence and the conditions for rational (jury) evaluation, Melbourne Univ. Law Rev. 39 (2015) 77.
[41] G. Edmond, B. Found, K. Martire, K. Ballantyne, D. Hamer, R. Searston, J. Tangen, Model forensic science, Aust. J. Forensic Sci. 48 (5) (2016) 496–537.
[42] Supreme Court of Victoria, Practice Note SC CR 3 Expert Evidence in Criminal Trials, 2017.
[43] B.L. Garrett, G. Mitchell, The proficiency of experts, Univ. Pennsylvania Law Rev. 166 (2018) 901–960.
[44] L. Wilson-Wilde, S. Smith, E. Bruenisholz, The analysis of Australian proficiency test data over a ten-year period, Forensic Sci. Policy Manage.: Int. J. 8 (1–2) (2017) 55–63.
[45] K.A. Neuendorf, *The Content Analysis Guidebook*, 2nd ed, Sage Publications, United States, 2016.
[46] K. Krippendorff, *Content Analysis: an Introduction to its Methodology*, 3rd ed, Sage Publications, United States, 2013.
[47] D. Riffe, S. Lacy, F. Fico, *Analyzing Media Messages: Using Quantitative Content Analysis in Research*, 3rd ed., Routledge, New York, NY, 2014.
[48] Krippendorff, K., Computing Krippendorff's Alpha-Reliability, 2011. Retrieved from http://repository.upenn.edu/asc_papers/43.
[49] K. Krippendorff, Content Analysis: An Introduction to its Methodology, Sage, Thousand Oaks, California, 2004.
[50] B.A. Spellman, Communicating forensic evidence: Lessons from psychological science, Seton Hall Law Rev. 48 (3) (2017) 827–840.
[51] American Association for the Advancement of Science [AAAS], Forensic Science Assessments: A Quality and Gap Analysis- Latent Fingerprint Examination, 2017. doi:10.1126/srhrl.aag2874.
[52] C.E. Berger, K. Slooten, The LR does not exist, Sci. Justice 56 (5) (2016) 388–391, https://doi.org/10.1016/j.scijus.2016.06.005.
[53] K.A. Martire, B. Growns, D.J. Navarro, What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts, Psychon. Bull. Rev. 25 (6) (2018) 2346–2355, https://doi.org/10.3758/s13423-018-1448-3.
[54] K.A. Martire, Clear communication through clear purpose: understanding statistical statements made by forensic scientists, Aust. J. Forensic Sci. 1–9 (2018), https://doi.org/10.1080/00450618.2018.1439101.
[55] G. Gigerenzer, A. Edwards, Simple tools for understanding risks: from innumeracy to insight, Br. Med. J. 327 (7417) (2003) 741–744, https://doi.org/10.1136/bmj.327.7417.741.
[56] G. Gigerenzer, R. Hertwig, E. van den Broek, B. Fasolo, K.V. Katsikopoulos, "A 30% chance of rain tomorrow": how does the public understand probabilistic weather forecasts? Risk Anal. 25 (3) (2005) 623–629, https://doi.org/10.1111/j.1539-6924.2005.00608.x.
[57] D.A. Nance, S.B. Morris, Juror understanding of DNA evidence: an empirical assessment of presentation formats for trace evidence with a relatively small random-match probability, J. Legal Stud. 34 (2) (2005) 395–444, https://doi.org/10.1086/428020.
[58] 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.
[59] K.A. Martire, I. Watkins, Perception problems of the verbal scale: a reanalysis and application of a membership function approach, Sci. Justice 55 (4) (2015) 264–273, https://doi.org/10.1016/j.scijus.2015.01.002.
[60] G. Mitchell, B.L. Garrett, The impact of proficiency testing information and error aversions on the weight given to fingerprint evidence, Behav. Sci. Law 37 (2) (2019) 195–210.
[61] J.J. Koehler, How trial judges should think about forensic science evidence, Judicature 102 (2018) 28–38.
[62] G. Edmond, K. Martire, M. San Roque, Unsound law: Issues with (expert) voice comparison evidence, Melbourne Univ. Law Rev. 35 (2011) 52–112.