2012-08-16

Response to: DR AS 5388.3 Forensic analysis - Part 3 - Interpretation

Response prepared by:

Dr Geoffrey Stewart Morrison

Director
Forensic Voice Comparison Laboratory
School of Electrical Engineering & Telecommunications
University of New South Wales
Sydney, New South Wales
Australia

The following individuals are in substantial agreement with and endorse this response. Some provided feedback on earlier drafts of this response, which was considered in the formulation of the final version. They are listed in the order in which I received their endorsements.

Dr Ian W Evett

Forensic Statistician
Principal Forensic Services Ltd
London, England
United Kingdom of Great Britain and Northern Ireland

Dr S M Willis

Director
Forensic Science Laboratory
Garda Headquarters
Dublin
Republic of Ireland

Prof Christophe Champod

Professor of Forensic Science
Faculty of Law and Criminal Justice
University of Lausanne
Lausanne
Switzerland

Dr Catalin Grigoras

Director
National Center for Media Forensics
University of Colorado Denver
Denver, Colorado
United States of America

Mr Jonas Lindh

Research Engineer
University of Gothenburg
Gothenburg
Sweden

CEO and Forensic Phonetic Consultant
Voxalys AB
Gothenburg
Sweden

Prof Norman Fenton

Director
Risk Information Management Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, England
United Kingdom of Great Britain and Northern Ireland

Dr Amanda Hepler

Senior Analyst
Innovative Decisions Inc
Vienna, Virginia
United States of America

Prof Charles E H Berger

Professor of Criminalistics
Faculty of Law
Leiden University
Leiden
The Netherlands

Dr John S Buckleton

Principal Scientist
Institute of Environmental Science and Research
Auckland
New Zealand

Prof William C Thompson

Professor
Department of Criminology, Law & Society and School of Law
University of California, Irvine
Irvine, California
United States of America

Prof Joaquín González-Rodríguez

Professor
ATVS - Biometric Recognition Research Group
ICFS-UAM Forensic Science Research Institute
Universidad Autonoma de Madrid
Madrid
Spain

Prof Cedric Neumann

Professor
Statistics Department
Eberly College of Science
The Pennsylvania State University
University Park, Pennsylvania
United States of America

Dr James M Curran

Associate Professor
Department of Statistics
University of Auckland
Auckland
New Zealand

Prof Cuiling Zhang

Director of Forensic Speech Science Section
Department of Forensic Science & Technology
China Criminal Police University
Shenyang
China

Visiting Professorial Fellow
Forensic Voice Comparison Laboratory
School of Electrical Engineering & Telecommunications
University of New South Wales
Sydney, New South Wales
Australia

Prof Colin Aitken

Professor of Forensic Statistics
The University of Edinburgh
Edinburgh, Scotland
United Kingdom of Great Britain and Northern Ireland

Dr Daniel Ramos

Assistant Professor
ATVS - Biometric Recognition Research Group
ICFS-UAM Forensic Science Research Institute
Universidad Autonoma de Madrid
Madrid
Spain

Lt Col José Juan Lucena Molina

Servicio de Criminalística
Dirección General de la Guardia Civil
Madrid
Spain

Dr Graham Jackson

Consultant Forensic Scientist
Advance Forensic Science
Dundee, Scotland
United Kingdom of Great Britain and Northern Ireland

Visiting Professor of Forensic Science
University of Abertay
Dundee, Scotland
United Kingdom of Great Britain and Northern Ireland

Dr Didier Meuwly

Principal Scientist
Netherlands Forensic Institute
The Hague
The Netherlands

Mr Bernard Robertson

Editor
New Zealand Law Journal
New Zealand

Dr G A Vignaux

Emeritus Professor of Operations Research
Victoria University of Wellington
Wellington
New Zealand

The endorsers and I are acting in our personal capacities, and the opinions expressed herein do not necessarily represent the official policies of the organisations with which we are affiliated.

Introduction

The draft standards appear to have essentially ignored the progress made in development of theory and practical implementation in the field of interpretation of forensic evidence over approximately the last twenty years, including in the interpretation of DNA evidence.

If the draft standards were adopted substantially as is, they would falsely legitimise current bad practice. This would be a major impediment to improving how forensic evidence is interpreted and reported in Australia. If bad practice were challenged in court or elsewhere, the offending party could claim that they are following the standards and thereby obtain an undeserved imprimatur, and perhaps avoid further scrutiny. See Lucena-Molina et al (2012) on the dangers of this already occurring in Spain because of legislative changes regarding the status of forensic reports. Ultimately, it would be better to have no published standards than to have bad published standards.

Rather than attempt to fix the current draft it would be more efficient to begin the process afresh inviting internationally acknowledged experts in the field of interpretation of forensic evidence to be involved in the drafting process from the very beginning, and paying particular attention to their advice.

In my opinion, the following would be essential components of standards for the evaluation and interpretation of forensic evidence when the forensic scientist's purpose is to assist a trier of fact in judicial proceedings.

1. Use of the likelihood-ratio framework

2. Testing of validity and reliability under conditions reflecting those of the case under investigation

3. Use of quantitative measurements, databases reflecting the relevant population, and statistical models

Each of these is outlined below.

1. Use of the likelihood-ratio framework

In Evett et al (2011) 31 leading experts in the interpretation of forensic evidence signed a statement to the effect that the likelihood-ratio framework is the logically most appropriate framework for the evaluation of forensic evidence. This statement was also endorsed by the Board of the European Network of Forensic Science Institutes (ENFSI), representing 58 laboratories in 33 countries. The likelihood-ratio framework has also been adopted by the Association of Forensic Science Providers in the UK and Republic of Ireland (AFSP, 2009), the former UK Forensic Science Service (Cook et al, 1998), and the Netherlands Forensic Institute (Berger, 2010), was argued for by the UK Forensic Science Regulator in R v T ([2010] EWCA Crim 2439, [2011] 1 Cr App R 9 at [77]), and recommended by the Report on Expert Evidence in Criminal Proceedings in England and Wales (Law Commission of England & Wales, 2011, §7.21(2c)). The statement has also been followed up and expanded upon by a number of articles in refereed forensic-science, general-science, and law journals, including Berger at al (2011), Robertson et al (2011), Redmayne et al (2011), Fenton (2011), Morrison (2012), Nordgaard & Rasmusson (2012), and Thompson (2012).

In the mid 1990's the likelihood-ratio framework was adopted as standard for the forensic comparison of DNA profiles (Foreman et al, 2003), and it is being gradually adopted in other branches of forensic science, including forensic voice comparison (Morrison, 2009), handwriting comparison (e.g., Hepler et al, 2012), fingerprint-fingermark comparison (e.g., Neumann et al, 2012), shoe marks (e.g., Skerrett et al, 2011), questioned documents (e.g., Neumann & Margot, 2009), firearms and toolmarks (e.g., Champod et al, 2003), glass (e.g., Curran et al, 2000), paint (e.g., McDermott et al, 1999), and fibres (e.g., Champod & Taroni, 1997).

In-depth descriptions of the likelihood-ratio framework can be found in numerous books and articles, including Robertson & Vignaux (1995), Aitken & Taroni (2004), Balding (2005), Buckleton (2005), Morrison (2010), and Aitken et al (2010), the latter prepared under the auspices of the Royal Statistical Society's Working Group on Statistics and the Law (note also that the authors of the first and third through fifth of these are either from or are working in Australasia).

Buckleton (2005) provides a particularly lucid comparison of the likelihood-ratio framework and frequentist approaches (allowed in the draft standards), and comes out in favour of the likelihood-ratio framework. Foreman et al (2003) document how frequentist approaches were abandoned for the evaluation of DNA evidence and replaced by the likelihood-ratio framework. The standards should not endorse frequentist approaches to the interpretation of evidence, which are outdated and logically inappropriate.

The draft standards endorse the expression of the results of forensic evaluation in the form of posterior probabilities. A forensic scientist cannot logically express results in such terms because they would require assignment of prior probabilities which are the domain of the trier of fact. Also, when the trier of fact considers all the evidence there is no logically correct mechanism whereby posterior probabilities derived from different piece of evidence can be combined (other than by knowing the specific priors used for each piece of evidence and removing their effect so as to revert to a likelihood ratio for each piece of evidence). The standards should specifically state that forensic scientists should not attempt to calculate or assign posterior probabilities.

Standards for the interpretation of forensic evidence would need to include a description of the likelihood-ratio framework (this could be developed from and make reference to the existing literature), and guidance as to how to implement the framework. A summary of the case assessment and interpretation model (CAI, e.g., Cook et al, 1998; Jackson et al, 2006), of which the likelihood-ratio is a core component, would provide guidance to the forensic scientist in selecting appropriate prosecution and defence hypotheses, including whether it is appropriate for them to address source level or activity level hypotheses. Guidance as to how to select the relevant population as defined in the defence hypothesis should also be included (e.g., Robertson & Vignaux, 1995, ch. 3; Aitken & Taroni, 2004, p. 274–281; Champod et al, 2004; Lucy, 2005, p. 129–133; Morrison et al , 2012). It is important to understand that a likelihood ratio is a strength-of-evidence statement in answer to a specific question defined by the prosecution and defence hypotheses, and that value of a likelihood ratio cannot be interpreted without understanding those hypotheses.

The standards should specifically state that the likelihood-ratio framework is currently the only acceptable framework for the evaluation of the strength of forensic evidence.

2. Testing of validity and reliability under conditions reflecting those of the case under investigation

While concern about the logically correct framework for the evaluation of forensic evidence has been particularly strong in Europe and Australasia, concern about the validity and reliability of forensic science has been particularly strong in the United States. In 1993 the Daubert ruling (Daubert v Merrell Dow Pharmaceuticals (92-102) 509 US 579 [1993]) was particularly concerned that the "evidentiary reliability" of forensic techniques be demonstrated (the Court equated "evidentiary reliability" with "scientific validity") and set a new standard for the admissibility of forensic evidence (although in practice the ideals of the standard are not always met). The Law Commission of England & Wales (2011) has also recently recommended an admissibility standard based on "evidentiary reliability".

In 2009 the US National Research Council published a report to Congress on Strengthening Forensic Science in the United States (NRC, 2009), which found that:

"[S]ome forensic disciplines are supported by little rigorous systematic research to validate the discipline's basic premises and techniques. There is no evident reason why such research cannot be conducted" (p. 22).

"The development of scientific research, training, technology, and databases associated with DNA analysis have resulted from substantial and steady federal support for both academic research and programs employing techniques for DNA analysis. Similar support must be given to all credible forensic science disciplines if they are to achieve the degrees of reliability needed to serve the goals of justice." (p. 13)

The report urged that procedures be adopted which include "quantifiable measures of the reliability and accuracy of forensic analyses" (p. 23), "the reporting of a measurement with an interval that has a high probability of containing the true value" (p. 121), and "the conducting of validation studies of the performance of a forensic procedure" (p. 121)

I take the position that use of the logically correct framework for the evaluation of evidence is a necessary precursor to evaluating the validity and reliability of forensic science. I therefore disagree with the particular metric for measuring validity proposed in the NRC report, which depends on binary decisions based on posterior probabilities, and with the particular metric for measuring validity proposed in Appendix B3 of the draft standards, which is based on frequentist rejection or failure to reject the null hypothesis (note that even from a frequentist perspective the draft standards are incorrect in equating failure to reject the null hypothesis with support for the null hypothesis). Instead I recommend metrics of validity and reliability consistent with the rôle of the forensic scientist in the likelihood-ratio framework, i.e., metrics based on the continuously-valued likelihood-ratio output of a forensic-evaluation system (Morrison, 2011) (a system includes the types of measurements made, techniques used to make those measurements, databases used, statistical models used, and all elements involving input from human experts).

I also take the position that testing of the validity and reliability of a forensic-evaluation system is not generally something which can be done once and then extrapolated to multiple different casework conditions. Rather, to the extent that the conditions of each case are different from the conditions of other cases, the validity and reliability of a forensic-evaluation system must be assessed under conditions reflecting as closely as possible those of the particular case under investigation, including using test data taken from the relevant population (Morrison, 2011; Morrison et al, 2012). Tests of the performance of a system under one set of conditions are not necessarily informative as to the performance of that system under other conditions.

The standards should specify that the validity and reliability of the forensic-evaluation system be assessed under conditions reflecting as closely as possible those of the case under investigation and using metrics consistent with the rôle of the forensic scientists within the likelihood-ratio framework. Also, that the results of these evaluations be provided to the judge when considering admissibility, and to the trier of fact when considering the strength-of-evidence statement derived via the use of the forensic-evaluation system. Systems which have not been so tested should not be used for the evaluation of forensic evidence.

3. Use of quantitative measurements, databases reflecting the relevant population, and statistical models

Testing of validity and reliability is a hallmark of modern scientific practice, the use of data, measurements, and statistical models are also hallmarks of modern scientific practice, they provide transparency and facilitate replicability. A forensic report (or case notes accessible to the court), like a scientific research report, should contain a sufficient description of materials and methodology that another suitably qualified scientist or team of scientists in a suitably equipped laboratory can replicate the analysis and (under normal circumstances) arrive at substantially the same result.

If there is substantial disagreement between the output of two forensic-evaluation systems each using data, measurements, and statistical models, then the source of this disagreement can be traced through any differences in choice of the type of measurements made, measurement techniques, choice of database reflecting the relevant population, choice of statistical modelling techniques, etc. There is no such thing as absolute objectivity, and examination of the choices listed above may reveal differences in opinion as to appropriate databases, measurement, modelling techniques, etc. Such an examination would potentially reveal very specific underlying causes of differences in system output, which could be debated by forensic scientists and potentially resolved pre trial, or which could be disclosed, explained, and considered in the trial process.

The transparency and replicability of approaches based on data, measurements, and statistical models contrasts with approaches based entirely or primarily on the experience of a forensic examiner. Subjective experience-based decisions are not easily replicated nor as open to the tracing and potential resolution of underlying sources of disagreement, and are more susceptible to human bias (a particular concern of the NRC report).

I consider the use of the likelihood-ratio framework and testing of validity and reliability under conditions reflecting those of the case at trial to be obligatory. In contrast, I consider the use of quantitative measurements, databases reflecting the relevant population, and statistical models to be highly preferable. The reason for this is that irrespective of the approach taken to arrive at a likelihood ratio, the validity and reliability of the forensic-evaluation system should be assessed under conditions reflecting those of the case at trial, and whichever system is found to have the highest degree of validity and reliability should be employed irrespective of its architecture (when systems are compared they must all be assessed on the same set of test data). Such testing provides a principled way of comparing experience-based systems with data-and-statistical-model based systems assuming the experience-based practitioner is willing and able to undergo the testing regime. Refusal or inability to undergo testing should disqualify the system from being used for the evaluation of forensic evidence, especially if a testable alternative system is available. Note also that experience-based judgments expressed on an ordinal or continuous scale can be treated as data and converted to likelihood ratios using statistical models, assuming the experience-based practitioner is willing and able to perform sufficient tests to provide training data for the models (Lindh & Morrison,

2010; Ramos et al, 2010).

The standards should express a strong preference for systems using quantitative measurements, databases reflecting the relevant population, and statistical models.

References

Aitken CGG, Roberts P, Jackson G (2010). *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*. London, UK: Royal Statistical Society. http://www.rss.org.uk/site/cms/contentviewarticle.asp?article=1132

Aitken CGG, Taroni F (2004). *Statistics and the Evaluation of Forensic Evidence for Forensic Scientist*, 2nd ed. Chichester, UK: Wiley.

Association of Forensic Science Providers (2009). "Standards for the Formulation of Evaluative Forensic Science Expert Opinion". *Science & Justice*, 49, 161–164. doi:10.1016/j.scijus.2009.07.004

Balding DJ (2005). *Weight-of-Evidence for Forensic DNA Profiles*. Chichester, UK: Wiley.

Berger CEH (2010). "Criminalistiek is terugredeneren" [Criminalistics is reasoning backwards]. *Nederlands Juristenblad*, 784–789.

Berger CEH, Buckleton J, Champod C, Evett IW, Jackson G (2011). "Evidence evaluation: A response to the Court of Appeal judgment in R v T". *Science & Justice*, 51, 43–49. doi:10.1016/j.scijus.2011.03.005

Buckleton J (2005). "A framework for interpreting evidence". In Buckleton J, Triggs CM, Walsh SJ (Eds.), *Forensic DNA Evidence Interpretation* (pp. 27–63). Boca Raton, FL: CRC.

Champod C, Baldwin D, Taroni F, Buckleton JS (2003). "Firearms and tool marks identification: The Bayesian approach. *Association of Firearm and Toolmark Examiners Journal*, 35, 307–316.

Champod C, Evett IW, Jackson G (2004). "Establishing the most appropriate databases for addressing source level propositions". *Science & Justice*, 44,153–164. doi:10.1016/S1355-0306(04)71708-6

Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA (1998). "A hierarchy of propositions: Deciding which level to address in casework". *Science & Justice*, 38, 231–239. doi:10.1016/S1355-0306(98)72117-3

Champod C, Taroni F (1997). "Bayesian framework for the evaluation of fibre transfer evidence". *Science & Justice*, 37, 75–83. doi:0.1016/S1355-0306(97)72151-8

Curran JM, Hicks-Champod TN, Buckleton JS (2000). *Forensic Interpretation of Glass Evidence*. Boca Raton, FL: CRC Press.

Evett IW, and other signatories (2011). “Expressing evaluative opinions: A position statement”. *Science & Justice*, 51, 1–2. doi:10.1016/j.scijus.2011.01.002

Fenton N (2011). “Improve statistics in court”. *Nature*, Vol. 479, 3 November, 36–37.

Foreman LA, Champod C, Evett IW, Lambert JA, Pope S (2003).”Interpreting DNA evidence: A review”. *International Statistics Journal*, 71, 473–495. doi:10.1111/j.1751-5823.2003.tb00207.x

Hepler AB, Saunders CP, Davis LJ, Buscaglia J (2012). “Score-based likelihood ratios for handwriting evidence”. *Forensic Science International*, 219, 129–140. doi:10.1016/j.forsciint.2011.12.009

Jackson G, Jones S, Booth G, Champod C, Evett IW (2006). “The nature of forensic science opinion: A possible framework to guide thinking and practice in investigation and in court proceedings”. *Science & Justice*, 46, 33–44. doi:10.1016/S1355-0306(06)71565-9

Law Commission of England & Wales (2011). *Expert evidence in criminal proceedings in England and Wales*. Report No. 235. London UK: The Stationery Office. http://www.lawcom.gov.uk/expert_evidence.htm

Lindh J, Morrison GS (2011). “Forensic voice comparison by humans and machine: Forensic voice comparison on a small database of Swedish voice recordings”. In Lee W-S, Zee E (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China* (pp. 1254–1257). Hong Kong: Organizers of ICPhS XVII at the Department of Chinese, Translation and Linguistics, City University of Hong Kong.

Lucena-Molina JJ, Pardo-Iranzo V, González-Rodríguez J (2012). “Weakening forensic science in Spain: From expert testimony to documentary evidence”. *Journal of Forensic Sciences*, 57, 952–963. doi:10.1111/j.1556-4029.2011.02041.x

Lucy D (2005). *Introduction to Statistics for Forensic Scientists*. Chichester, UK: Wiley.

McDermott SD, Willis SM, McCullough JP (1999). “The evidential value of paint. Part II: a Bayesian approach”. *Journal of Forensic Science*, 44, 263–269.

Morrison GS (2009). “Forensic voice comparison and the paradigm shift”. *Science & Justice*, 49, 298–308. doi:10.1016/j.scijus.2009.09.002

Morrison GS (2010). “Forensic voice comparison”. In Freckelton I, Selby H (Eds.), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson Reuters. http://expert-evidence.forensic-voice-comparison.net/

Morrison GS (2011). “Measuring the validity and reliability of forensic likelihood-ratio systems”. *Science & Justice*, 51, 91–98. doi:10.1016/j.scijus.2011.03.002

Morrison GS (2012). “The likelihood-ratio framework and forensic evidence in court: A response to R v T”. *International Journal of Evidence and Proof*, 16, 1–29. doi:10.1350/ijep.2012.16.1.390

Morrison GS, Ochoa F, Thiruvaran T (2012). “Database selection for forensic voice comparison”. In *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop, Singapore* (pp. 62–77). International Speech Communication Association.

National Research Council, (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: National Academies Press. http://www.nap.edu/catalog.php?record_id=12589

Neumann C, Evett IW, Skerrett J (2012). “Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm”. *Journal of the Royal Statistical Society A*, 175, 371–415. doi:10.1111/j.1467-985X.2011.01027.x

Neumann C, Margot P (2009). “New perspectives in the use of ink evidence in forensic science - Part III: Operational applications and evaluation”. *Forensic Science International*, 192, 29–42. doi:10.1016/j.forsciint.2009.07.013

Nordgaard A, Rasmusson B (2012). “The likelihood ratio as value of evidence- More than a question of numbers”. *Law, Probability & Risk*, online. doi:10.1093/lpr/mgs019

Ramos D, Franco-Pedroso J, González-Rodríguez J (2011) “Calibration and weight of the evidence by human listeners: The ATVS-UAM submission to NIST Human-Aided Speaker Recognition 2010”. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), Prague, Czech Republic* (pp. 5908–5911).

Redmayne M, Roberts P, Aitken CGG, Jackson G (2011). “Forensic science evidence in question”. *Criminal Law Review*, 5, 347–356.

Robertson B, Vignaux GA (1995). *Interpreting Evidence*. Chichester, UK: Wiley.

Robertson B, Vignaux GA, Berger CEH (2011) “Extending the confusion about Bayes”. *Modern Law Review*, 74, 444–455. doi:10.1111/j.1468-2230.2011.00857.x

Skerrett J, Neumann C, Mateos-García I (2011). “A Bayesian approach for interpreting shoemark evidence in forensic casework: Accounting for wear features”. *Forensic Science International*, 210, 26–30. doi:10.1016/j.forsciint.2011.01.030

Thompson, W.C. (2012). “Bad cases make bad law: Reactions to R v T”. *Law, Probability & Risk*, online. doi:10.1093/lpr/mgs020