From: [redacted]
To: [redacted]
Subject: bitemarks
Date: Friday, December 2, 2016 8:27:09 PM

Dear Dr. Lander,

I have just been notified by the AAFS that your Council is requesting additional information on certain pattern-matching disciplines used in US criminal courts. This attachment is a recently published journal article that, in my opinion, is material to your continuing review of bitemark analysis involving injury patterns on human skin. It may prove to be a counter-point to any other material your request for bitemark validation data may produce. Please pardon me if your subcommittee has already seen the material.

https://academic.oup.com/jlb/article/2544494/Forensic-bitemark-identification-weak-foundations

Regards,

C. Michael Bowers DDS JD
Fellow, AAFS, Odontology
Author: Forensic Testimony: Science, Law and Expert Evidence

Salus populi suprema lex

"Public safety is the highest law"

From: [redacted]
To:
Cc: [redacted]
Subject: Re: Invitation to Provide Follow-up Information to PCAST Regarding its Forens...
Date: Sunday, December 4, 2016 6:50:45 PM

Dear Dr. Lander:

Thank you and colleagues for sharing this important information. I do not have additional reports or references to suggest. However, in my experience, sample contamination due to problems in chain-of-custody and mishandling can cause unrecognized errors in DNA and other analyses. I once demonstrated this point in a presentation I gave at an AAFS meeting on the mathematics and use of PCR.

It is suggested that a special research program be established to examine fundamental aspects of contamination that impact on the areas in the report. I think this can be an overriding issue to the many topics included in the report provided to AAFS members and others.

Many thanks.

Sincerely yours,

Walter

Walter E. Goldstein, Ph.D., PE
President
Goldstein Consulting Company
120 Emerald Forest Street, Unit #103
Las Vegas, Nevada 89145-3987
www.goldconsul.com
[redacted]
www.ivrbc.info

From: Bieber, Frederick R.,Ph.D.
To: [redacted]
Cc:
Subject: forensic DNA mixtures
Date: Monday, December 5, 2016 4:25:28 PM

Re: DNA mixtures

Dear Dr. Lander and members of PCAST:

Within the limits of time, I offer your team some brief comments in response to your recent email about accuracy of various forensic methods currently in use. There is, as you know, a large body of published literature on DNA mixture interpretation and validation using new software tools. The large amount of developmental validation work on DNA mixtures, performed internally by forensic laboratories, is not typically published as it is not considered "original research". The FBI DNA Advisory Board, NRCII, and other international groups (e.g., ISFG) have carefully considered mixtures as well. Whether individual laboratories follow appropriate recommended practices in individual cases is a different question that needs resolution at the local level.

Interpretation of evidentiary DNA mixtures is of great import to the forensic community. A general comment, worthy of note, applies to interpretation of forensic DNA mixtures when different swabs are joined together for DNA extraction, subsequent PCR amplification, and STR analysis. First, while merging swabs from different sources, or from different swabs of the same item, a mixture can be created where none actually existed. Second, the different swabs could contain different amounts of DNA from one or more individuals, leading to confusion about major/minor contributors in any mixture that is detected. Third, an individual could be falsely implicated as having handled/possessed an item when he did not, if his DNA was present on only one of the components of the mixture which was created by this common laboratory practice. This comment, while important in a general sense, is not relevant to all cases.

In your consideration of forensic DNA mixture interpretation, please be reminded that the most difficult/challenging forensic DNA mixtures are NOT those in which there are unequal contributions from 2 or 3 contributors, but rather mixtures with equivalent, or nearly equivalent, contributions from 2 (or 3, or 4) contributors. For example, a 10:4:1 mixture would often be easier to interpret (using CPI or newly developed software tools) than a 3:2:1 or a 2:1:1 mixture, assuming of course that sufficient DNA is present in the minor component to meet validation thresholds for allele calling.

All forensic mixtures are NOT created equal, as they range from the straightforward and simple to the most challenging and irresolvable. Even the new probabilistic genotyping methods can present practical challenges, to wit:

1. The two competing hypotheses (e.g., the numerator and the denominator in the likelihood ratios) are chosen by the laboratory, often without consultation with the advocates (i.e., the prosecutor and defense counsel) and without knowledge of the so-called fact pattern in the instant case.
2. The probabilistic genotyping software requires an "assumption" of the number of

contributors for each "run".

3. The Markov chain Monte Carlo (MCMC) methods used by some of the new software will produce a slightly different output each time the same mixture is input,,,which could lead to downstream confusion in the courtroom....this will need to be addressed in teaching to judges, attorneys, juries, et al.

If I have more time before your December 14th deadline I will write more.
Best regards,

Fred B.

Frederick R. Bieber
Medical Geneticist, Brigham and Women's Hospital

Associate Professor of Pathology
Harvard Medical School

[redacted]

From: [redacted]
To: [redacted]
Cc: [redacted]
Subject: feedback on PCAST Forensic Science Report
Date: Wednesday, December 7, 2016 11:59:47 AM
Attachments: pcast_forensic_science_report_final.pdf

To Whom It May Concern:

There are two instances where the term "voiceprint" is used in this document. The term "voiceprint" is problematic and is not used the by scientific forensic voice comparison community. We are trying to extinguish the use of this word which equates the human voice as a biometric with fingerprints. Speaking scientifically, fingerprint patterns do not change from moment to moment or day to day; and very little within a person's lifetime. The human voice, however, is a highly dynamic signal which undergoes changes throughout the day, many times within a person's life, and depending on mood and meaning of communication. Equating fingerprints to the voice is dangerous especially as so-called private experts working on forensic voice comparison cases convince juries, etc. of that this fallacy is true.

I would suggest changing language in the document as follows:

- Page 46 "voice samples"
- Page 64 "voice comparison"

Thanks, Jeff

Jeff M. Smith
Associate Director
National Center for Media Forensics
College of Arts & Media
University of Colorado Denver
[redacted]

From: [redacted]
To:
Subject: INTERPOL literature reviews by forensic discipline
Date: Thursday, December 8, 2016 4:12:05 PM

Every three years forensic experts from around the world gather at INTERPOL headquarters in Lyon, France to discuss the literature in various forensic science disciplines. Unfortunately, until recently these literature listings were not available on a public website. Review articles from the last two INTERPOL meetings (held in October 2013 and October 2016) can now be downloaded from their website: https://www.interpol.int/INTERPOL-expertise/Forensics/Forensic-Symposium.

My summary of their contents is included below along with direct links to the full documents covering literature from 2010-2016. While many articles are cited in these reviews, there is typically not a lot of analysis to demonstrate how these articles may or may not establish any kind of foundational validity or assist in estimating the accuracy of provided methods.

- **John Butler, National Institute of Standards and Technology**

The 2013-**2016** INTERPOL Literature Summary contains 4891 references to the following disciplines:

https://www.interpol.int/content/download/33314/426506/version/1/file/INTERPOL%2018th%20IFSMS%20Review%20Papers.pdf (769 page, 8.5 MB pdf file)

Topic	Authors (affiliations)	# References
Firearms	Erwin J.A.T. Mattijseen (Netherlands Forensic Institute)	179
Forensic Geosciences	Lorna Dawson (James Hutton Institute, Aberdeen, UK)	245
Gun Shot Residue	Sébastien Charles, Bart Nys, Nadia Geusens (INCC-NICC Brussels, Belgium)	77
Marks	Martin Baiker (Netherlands Forensic Institute)	104
Paint and Glass	Jose Almirall (Florida International University, USA)	102
Fibers and Textiles	Laurent Lepot, Kris De Wael, Kyra Lunstroot (INCC-NICC Brussels, Belgium)	92
Fire Investigation & Debris Analysis	Eric Stauffer (University of Lausanne, Switzerland)	194
Explosives	Douglas J. Klapec and Greg Czarnopys (ATF Laboratory, USA)	646
Drugs	Robert F.X. Klein (Drug Enforcement Administration Laboratory, USA)	1434
Toxicology	Wing-man Lee, Kwok-leung Dao, Wing-sum Chan, Tai-wai Wong, Chi-wai Hung, Yau-Nga Wong, Lok-hang Tong, Kit-mai Fung, Chung-wing Leung (Hong Kong Government Laboratory, China)	600
Audio	Catalin Grigoras, Andrzej Drygajlo, Jeff M. Smith (University of Colorado-Denver, USA and École Polytechnique Fédérale de Lausanne, Switzerland)	88
Video and Imaging	Arnout Ruifrok, Zeno Geradts, (Netherlands Forensic Institute)	108
Digital Evidence	Paul Reedy (Department of Forensic Science, District of Columbia, USA)	100
Fingermarks and Other Impressions	Andy Bécue and Christophe Champod (University of Lausanne, Switzerland)	536
DNA and Biological Evidence	Francois-Xavier Laurent and Laurent Pene (Institut National de Police Scientifique, Cedex, France)	75
Questioned Documents	Julien Retailleau (IRCGN, Pontoise, France)	255
Forensic Science Management	William P. McAndrew (Gannon University, Erie, PA, USA) and Max M. Houck (Forensic & Intelligence Services LLC, USA)	56

The 2010-**2013** INTERPOL Literature Summary contains 4832 references to the following disciplines:

https://www.interpol.int/content/download/21910/206602/version/1/file/IFSMSReviewPapers2013.pdf (928 pages; 3.9 MB pdf file)

Topic	Authors (affiliations)	# references
Firearms	Erwin J.A.T. Mattijseen (Netherlands Forensic Institute)	159
Gun Shot Residue	Sébastien Charles and Bart Nys (INCC-NICC Brussels, Belgium)	49
Toolmarks	Nadav Levin (Israel National Police)	189
Paint	Laetitia Heudt, Marc Lannoy, Gilbert De Roy, Laurent Kohler (INCC-NICC Brussels, Belgium)	201
Fibers and Textiles	Ray Palmer (Northumbria University, UK)	68
Forensic Geology	Ritsuko Sugita, Hiromi Itamiya, Hirofumi Fukushima (National Research Institute of Police Science, Japan)	221 cited but only 102 references listed
Arson & Fire Debris Analysis	Niina Viitala and Mika Hyyppä (National Bureau of Investigation, Finland)	157 cited but only 140 references listed
Explosives & Explosive Residues	Douglas J. Klapec and Greg Czarnopys (Bureau of Alcohol, Tobacco, Firearms and Explosives, USA)	1341
Drug Evidence	Jeffrey H. Comperio and Robert F.X. Klein (Drug Enforcement Administration, USA)	668
Toxicology	Wai-ming Tam, Lai-chu Chim, Wing-sum Chan, Tai-wai Wong, Kit-mai Fung, Wing-cheong Wong, Wai-kit Lee, Wing-sze Lee, Kit-man Fan (Hong Kong Government Laboratory)	324
Forensic Audio Analysis	Catalin Grigoras, Jeff M. Smith, Geoffrey Stewart Morrison, Ewald Enzinger (University of Colorado-Denver, USA and University of New South Wales, Australia)	133
Forensic Video Analysis	Matthew E. Graves (United States Army Criminal Investigation Laboratory)	31
Imaging	Arnout Ruifrok, Zeno Geradts, Jerrien Bijhold (Netherlands Forensic Institute)	256
Digital Evidence	Paul Reedy and Jaime Buzzeo (Department of Forensic Science, District of Columbia and A.I. Solutions at NASA Headquarters, USA)	190
Fingermarks and Other Impressions	Nicole Egli, Sébastien Moret, Andy Bécue, Christophe Champod (University of Lausanne, Switzerland)	472
Body Fluid Identification and DNA Typing in Forensic Biology	Christine Jolicoeur (Ministry of Public Security, Québec, Canada)	114
Questioned Documents	Franck Partouche (IRCGN, Rosny Sous Bois, France)	275
Forensic Science Management	Max M. Houck, Melissa Porter, Bronwen Davies (Department of Forensic Sciences and George Washington University, Washington, DC, USA)	120

From: [redacted]
To: [redacted]
Cc: [redacted]
Subject: Re: Footwear analysis
Date: Sunday, December 11, 2016 11:46:19 AM
Attachments: PIIS0379073816304455.pdf

Sorry, the paper by Dr. Speir can be found at :https://dx.doi.org/10.1016/j.forsciint.2016.06.012
Attached.

On Sun, Dec 11, 2016 at 5:22 PM, Yaron Shor <[redacted]> wrote:

Eric Lander
Co-Chair, PCAST

)

Hello Mr. Lander

I´m writing to you on behalf of Mr. Nadav Levin from Israel, and our research team that includes: Dr. Yoram Yekutieli, Sarena Wiesner, Tsadok Tsach and myself.

Our research "Expert Assisting Computerized System for Evaluating the Degree of Certainty in 2D Shoeprints" was posted in the NIJ site after peer review by three reviewers. It might not be a scientific journal, and the paper about the research is on the way, but we feel that it will be beneficial to the field if the results will be mentioned in the PCAST report. The research was thorough and provided good scientific foundation to the accidental high value and discriminating opportunities. The big database of accidentals (RACs) and the algorithms developed in the project, show how a reliable estimation about the rarity of accidentals on shoe sole can be found and be shown.

The report was posted to NCJRS.gov on November 15th. the link below:

https://www.ncjrs.gov/pdffiles1/nij/grants/250336.pdf

Another paper that was written by our colleague, Jacqulin Speir titled "Quantitative Assessment of Similarity between Randomly Acquired Characteristics on High Quality Exemplars and Crime Scene Impressions via

Analysis of Feature Size and Shape". can be found at:
file:///C:/Users/Administrator/Downloads/PIIS0379073816304455.pdf

Dr. Speir also conducted a big research on RACs and their rarity, and her conclusions support the hypothesis that accidentals on shoe-sole can establish, in a scientific method, the base for the discriminating power shoe prints are supposed to establish.

We hope these papers will be mentioned in the next publication of PCAST.

Thanks

Yaron Shor

Israel Police HO,

DIFS, Toolmarks and Materials Lab.

Midwestern Association of Forensic Scientists Response to PCAST Report- December 12th, 2016

The President's Council of Advisors on Science and Technology (PCAST) has issued a Report to the President on Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, hereinafter referred to as PCAST Report. The PCAST report broadly defines Forensic Science, but quickly characterizes its goal to help close gaps "for the case of forensic 'feature-comparison' methods" and chooses to narrowly focus mainly on six forensic disciplines, (1) DNA analysis of single-source and simple-mixture samples, (2) DNA analysis of complex-mixture samples, (3) bite-marks, (4) latent fingerprints, (5) firearms identification, and (6) footwear analysis.

The Midwestern Association of Forensic Scientists (MAFS) would like to first acknowledge that the PCAST report makes some very good points that should be commended. Among many of the reports suggestions to the forensic science and legal communities are to strengthen "foundational validity", strengthen the measurement of uncertainty in conclusions, increase clarity in testimony, avoid scientifically indefensible claims, abate contextual bias and eliminate conformational bias. These suggestions are noteworthy and common goals of not only MAFS, but we would posit, all Forensic Scientists.

Where we believe the report fell short is to first not recognize the valuable contribution forensic science has provided to the criminal justice system. Many lives have been saved and victims vindicated by the work forensic scientists do every day. Secondly, as with any scientific profession we recognize that our science can always be improved, but for PCAST to broadly characterize it as lacking foundational or scientific validity is capricious. PCAST might not agree with the approach of much of the foundational research, but that does not discount that a considerable amount of research has been completed in each one of the disciplines targeted by this report.

More specifically, with regard to scientific research methods and validity, the PCAST report is replete with suggestions and assertions that "evaluations of validity and reliability must therefore be based on 'black-box studies'. Where we agree that black box studies can be useful or perhaps even essential, they certainly are not the only scientific way to ensure validity and reliability.

With regard to Proficiency testing, again MAFS would agree that proficiency testing can be improved and perhaps be more rigorous. While we can tacitly agree with the statement that "the only way to establish scientifically that an examiner is capable of applying a foundationally valid method is through appropriate empirical testing to measure how often the examiner gets the correct answer" we would disagree with the statement that "Such empirical testing is often referred to as 'proficiency testing'". MAFS does understand that PCAST and the Forensic Community may have differing opinions on the definition of Proficiency Test; however empirical testing is commonly known in the sciences as originating in or based on observation or experience and it is our opinion that the PCAST report is too quick to discount experience. Empirical testing of an examiner's abilities should be defined to include training exercises, mock cases, competency tests, validation studies, and case work experience, in addition to proficiency testing. These are all components of training and case work that are continually

performed in forensic laboratories. Experience and daily observation should not be discounted since the accumulation of training, testing and experience adds to the empirical knowledge of our particular discipline. Medical Doctors are not "proficiency tested" in their ability to diagnose the common cold, or even cancer; instead they rely on their experience and training. Moreover, where PCAST suggests "a forensic examiner's 'experience' from extensive casework is not informative—because the 'right answers' are not typically known'", we would posit that when a physician diagnoses the common cold, the right answer is "not known" either but he or she is relying heavily on experience (as well as other factors, seasonal, etc.).

MAFS strongly disagrees with the characterization that the "forensic community prefers that tests not be too challenging". This is an assertion based on one comment, and an opinion at that, by one president, from one test provider. This report is fraught with rhetoric about rigorous research, reproducibility, repeatability, etc. and it appears contradictory and careless to make such a hyperbolic statement based upon one person's opinion.

Regarding funding, we again agree with the PCAST report that in order to move forward with their suggestions to strengthen the science, more funding is needed; however, we disagree with their specific recommendations. PCAST recommends $4 million to support efforts to make methods "established as foundationally valid" which is predominately what the entire report is about; on this we marginally agree; however, then PCAST recommends "$10 million to support increased research activities in forensic science, including on complex DNA mixtures, latent fingerprints, voice/speaker recognition, and face/iris biometrics." With all due respect to the DNA community, a considerable amount of funding is already available through the current DNA Capacity Enhancement and Backlog Reduction (CEBR) Program. Perhaps the more egregious recommendation regarding funding is that of voice/speaker recognition, and face/iris biometrics. We find it puzzling to suggest funding for voice/speaker recognition, and face/iris biometrics when neither technology is mentioned as a concern in this report. We assert that the funding should go to research in the forensic science disciplines that are in fact the cause of the concern.

Lastly, with regard to many of the recommendations, for example establishing foundational validity and proficiency testing, PCAST recommends the involvement of independent scientists without direct forensic science experience or as stated in the report "which has no stake in the outcome". Where we would welcome more involvement from the academic community, statisticians, etc. and believe their involvement can only strengthen our science, we all know that science is about collaboration, discussion and debate. To not include practitioners in the discussion would be irresponsible.

Although we may disagree on many points that the PCAST report makes, the Midwestern Association of Forensic Scientists would like to thank PCAST for its work. We understand that the undertaking was immense and that we are not always going to agree, but we do stand united and ready to strengthen our science whenever and wherever the opportunity arises.

From: [redacted]
To: [redacted]
Subject: response to PCAST
Date: Tuesday, December 13, 2016 11:26:27 AM
Attachments: Koch response to PCAST.docx

I'd like to offer the attached review of hair analysis, foundational research, points of clarification, and reference list for PCAST to consider.

Sincerely,
Sandra Koch, F-ABC
PhD Candidate
Pennsylvania State University
Department of Anthropology

Forensic analysis of hair draws upon research from multiple disciplines: anthropological analysis of human variation, biology for aspects related to growth and development of the hair, the cosmetic industry for how artificial treatment affects hair structure, genetics for the search for links between genotype and phenotype, and forensic science for species identification and comparison between questioned and known samples. PCAST seemed to only look at forensic articles. Much of the foundational research, which is still valid, was conducted by anthropologists focused on identifying characteristics in hairs that would help differentiate ancestry such as macroscopic hair form (e.g. Turner et al. 1914; Trotter, 1938). Their early empirical studies focused on ancestry characterization through documentation and measurement of macroscopic hair form (e.g. Prunner-Bey, 1877; Saint-Hillaire, 1860; Trotter 1930, 1934, 1936, 1956). Mildred Trotter calculated the area of hair sections collected from numerous individuals within different population groups and separated by age and sex to determine the variation present in hair form by measuring the major and minor axis of the hair shaft cross sections (½ greatest diameter x ½ least diameter x π). Guilbeau-Frugier et al. (2006) found that hair form area calculations were able to differentiate the broad ancestry groups of European, Sub-Saharan African, and Asian populations but some populations overlapped to an extent that differentiation was not possible by calculation of hair shaft area, such as between European and North African populations. Hair classifications often fall into a continuum and are not easily differentiable into categories by discrete characteristics; however, forensic hair examiners are trained microscopists who consider all features when examining an evidential hair.

Research focused on specific layers of a hair has also added to our understanding of human variation and the features that can be useful in a microscopical comparison. Takahashi *et al.* (2015) noted differential compaction of the cuticle layers but that there were similar numbers of layers among different groups of peoples. Wynkoop (1929) distinguished 4 medullary types while Hausman (1930) reduced those types to the presence or absence of a medulla and attempted to determine if there was a correlation with hair diameter. Duggins and Trotter (1950) were unable to find a correlation between medullary types, the diameter of a hair, or the age of the individual. Banerjee's (1965) study showed that despite Hausman's (1930) early research which indicated a correlation, that the medullary structure does not have a clear relationship with hair form. Hrdy (1973) suggested the presence of a medulla and its thickness was correlated with overall hair diameter, most notably in individuals of Asian ancestry, but he also noted populations where this correlation was not held up, specifically between individuals from African and New Guinean individuals. From these works, the caution is to not emphasize the presence or absence of a specific feature as being associated with an ancestry group as there are degrees of human variation. Forensic microscopists take into account the hair form, the changes in diameter along the length of a hair, the thickness of the cuticle, etc. and use all the features visible in a hair to compare samples in order to decide whether to include or exclude a known sample as a potential source. This is not an identification and forensic hair examiners clearly state the limitations to microscopical examinations in their reports and testimony along with stating a need for the same hair to be analyzed by mitochondrial DNA (mtDNA) should identification be important to a case (if such statements are allowed by their agency).

Research by Kajiura(2006), Bryson et al. (2007), and Fugimoto et al. (2008) focused on potential links between hair form with genetic ancestry (African, Caucasian, and Asian) but more research is needed here. Hair examiners also rely on research from a variety of disciplines which seek to understand hair microstructure and changes to hair over the life of an individual, from death as well as environmental

changes such as deposition at a crime scene (Bryson, et al., 2009; Duggins 1954, Trotter and Duggins, 1948 and 1950; Koch et al., 2013; Hietpas et al., 2016; Roberts et al., 2016; Tridico et al. 2015; etc.). The PCAST committee appears to not have considered the foundational research from anthropology, or any of the additional research in biological or cosmetic research fields from which forensic hair examiners gain much knowledge.

I urge PCAST to use their platform not to denigrate specific scientific disciplines or methods (microscopical analysis is a valid comparison tool) but to urge laboratories who have gotten rid of microscopical analysis to bring it back as it is a useful and inexpensive screening technique with well-trained microscopists. Those who rush to send a hair for DNA analysis without first conducting a microscopical analysis destroy potentially valuable information – whether the hair was dyed, if there is evidence of decomposition, potential ancestry determinations, if a hair contains similar internal patterning of pigment granules, etc. This information along with the determination of whether a known sample can be included or not (microscopical hair analysis is not an identification but class type evidence) and then DNA analyses can and should be conducted. Not all evidential hairs have sufficient tissue for a nuclear DNA analysis and mitochondrial DNA is not a means of identification, but the combination of microscopy and mtDNA is stronger evidence about a hair and that combination should be pushed for in all labs and courts throughout the US. This combination was the point of the Houck and Budowle paper and one that PCAST appeared to misunderstand. The regional mtDNA labs, previously funded and trained by the FBI, should be brought back so hair evidence throughout the US can benefit from both microscopical and mitochondrial analysis prior to being presented in court. Microscopy can differentiate hairs between populations and individuals at a rate greater than chance alone as shown by the Houck and Budowle (2002) paper, Gaudette and Keeping, 1974, and Strauss, 1983 among others. Combining microscopy with mtDNA analysis increases the ability to further discriminate hairs among individuals. Stressing the need for the combination of techniques is where I believe PCAST can best use its influence.

I strongly caution the placement of any "science-based agency" being placed on a pedestal as being more scientifically reliable than bench scientists. NIST is not known to have scientists with a background in microscopical analysis of hairs. As the training and experience of a hair examiner is crucial to the reliability of this discipline, PCAST should strongly encourage NIST, NIJ, and academic researchers to include qualified hair examiners in their future research projects involving hair. This is critical to ensure any new research will be useful for forensic applications and be able to be easily adopted into state and local laboratory protocols for hair analysis. The guidance this committee gives can aid in promoting specific research aims; however, I'd caution this committee from wholesale denigration of a field as that serves no purpose. Forensic sciences should be subjected to rigorous cross examination in court and peer review for research.

Sincerely,
Sandra Koch
Fellow – American Board of Criminalistics (Hairs and Fibers)

Cited and suggested references for further reading by the committee:

Banerjee, A. R. (1965). On variation of human head hair: Hair form and medullation. *Zeitschrift für Morphologie und Anthropologie*, (H. 1), 56-69.

Bernard, B. A. (2003). Hair shape of curly hair. *Journal of the American Academy of Dermatology*, *48*(6), S120-S126.

Bryson, W. G., Harland, D. P., Caldwell, J. P., Vernon, J. A., Walls, R. J., Woods, J. L., ... & Koike, K. (2009). Cortical cell types and intermediate filament arrangements correlate with fiber curvature in Japanese human hair. *Journal of structural biology*, *166*(1), 46-58.

Das-Chaudhuri, A. B., & Chopra, V. P. (1984). Variation in hair histological variables: medulla and diameter. *Human heredity*, *34*(4), 217-221.

Duggins, O. H. (1954). Age changes in head hair from birth to maturity IV. Refractive indices and birefringence of the cuticle of hair of children. *American journal of physical anthropology*, *12*(1), 89-114.

Duggins, O. H., & Trotter, M. (1950). Age changes in head hair from birth to maturity. II. Medullation in hair of children. *American journal of physical anthropology*, *8*(3), 399-416.

Duggins, O. H., & Trotter, M. (1951). Changes in morphology of hair during childhood. *Annals of the New York Academy of Sciences*, *53*(3), 569-575.

Duggins, O. H., Trotter, M., & Coon, C. S. (1959). Hair from a Kadar woman of India. *American journal of physical anthropology*, *17*(2), 95-98.

Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L., ... & Morishita, Y. (2008). A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Human molecular genetics*, *17*(6), 835-843.

Fujimoto, A., Nishida, N., Kimura, R., Miyagawa, T., Yuliwulandari, R., Batubara, L., ... & Morishita, Y. (2009). FGFR2 is associated with hair thickness in Asian populations. *Journal of human genetics*, *54*(8), 461-465.

Gaudette, B. D., & Keeping, E. S. (1974). An attempt at determining probabilities in human scalp hair comparison. *Journal of Forensic Science*, *19*(3), 599-606.

Guilbeau-Frugier, C., Blanc, A., Crubezy, E., Delisle, M. B., Rouge, D., & Telmon, N. (2006). Hair morphology and anthropological applications. *American Journal of Human Biology*, *18*(6), 861-864.

Hausman, L. A. (1925). A comparative racial study of the structural elements of human head-hair. *The American Naturalist*, *59*(665), 529-538.

Hausman, L. A. (1925). The relationships of the microscopic structural characters of human head-hair. *American Journal of Physical Anthropology*, *8*(2), 173-177.

Hausman, L. A. (1927). The pigmentation of human head-hair. *American Naturalist*, 545-554.

Hausman, L. A. (1928). The pigment granules of human head hair: a comparative racial study. *American Journal of Physical Anthropology*, *12*(2), 273-283.

Hausman, L. A. (1934). Histological variability of human hair. *American Journal of Physical Anthropology*, *18*(3), 415-429.

Hall, R. (1950). Hair and some of its clinical significances. *The Ulster medical journal*, *19*(2), 133.

Hietpas, J., Buscaglia, J., Richard, A. H., Shaw, S., Castillo, H. S., & Donfack, J. (2016). Microscopical characterization of known postmortem root bands using light and scanning electron microscopy. *Forensic Science International, 267*, 7-15.

Houck, M. M., & Budowle, B. (2002). Correlation of microscopic and mitochondrial DNA hair comparisons. *Journal of forensic sciences, 47*(5), 964-967.

Hrdy, D. (1973). Quantitative hair form variation in seven populations. *American journal of physical anthropology, 39*(1), 7-17.

Kajiura, Y., Watanabe, S., Itou, T., Nakamura, K., Iida, A., Inoue, K., ... & Amemiya, Y. (2006). Structural analysis of human hair single fibres by scanning microbeam SAXS. *Journal of structural biology, 155*(3), 438-444.

Khumalo, N. P., Doe, P. T., Dawber, R. R., & Ferguson, D. J. P. (2000). What is normal black African hair? A light and scanning electron-microscopic study. *Journal of the American Academy of Dermatology, 43*(5), 814-820.

Khumalo, N. P., Stone, J., Gumedze, F., McGrath, E., Ngwanya, M. R., & de Berker, D. (2010). 'Relaxers' damage hair: Evidence from amino acid analysis. *Journal of the American Academy of Dermatology, 62*(3), 402-408.

Koch, S. L., Michaud, A. L., & Mikell, C. E. (2013). Taphonomy of hair—a study of postmortem root banding. *Journal of forensic sciences, 58*(s1).

Lasisi, T., Ito, S., Wakamatsu, K., & Shaw, C. N. (2016). Quantifying variation in human scalp hair fiber shape and pigmentation. *American journal of physical anthropology, 160*(2), 341-352.

Loussouarn, G., Garcel, A. L., Lozano, I., Collaudin, C., Porter, C., Panhard, S., ... & De La Mettrie, R. (2007). Worldwide diversity of hair curliness: a new method of assessment. *International journal of dermatology, 46*(s1), 2-6.

Nagase, S., Kajiura, Y., Mamada, A., Abe, H., Shibuichi, S., Satoh, N., ... & Amemiya, Y. (2009). Changes in structure and geometric properties of human hair by aging. *Journal of cosmetic science, 60*(6), 637.

Oien, C. T. (2009). Forensic hair comparison: background information for interpretation. *Forensic Science Communications–FBI, 1*(2).

Pruner-Bey, D. (1877). On human hair as a race character. *Journal of the Anthropological Institute of Great Britain and Ireland*, 71-92.

Roberts, K. A., Garcia, L. R., & De Forest, P. R. (2016). Proximal End Root Morphology Characteristics in Antemortem Anagen Head Hairs. *Journal of Forensic Sciences*.

Robertson, J. (1982). An appraisal of the use of microscopic data in the examination of human head hair. *Journal of the Forensic Science Society, 22*(4), 390-395.

Seibert, H. C., & Steggerda, M. (1942). The size and shape of human head hair along its shaft. *Journal of Heredity, 33*(8), 302-304.

Steggerda, M. (1940). Cross sections of human hair from four racial groups. *Journal of Heredity*, *31*(11), 474-476.

Steggerda, M. (1941). Change in hair colour with age. *Journal of Heredity*, *32*(11), 402-404.

Steggerda, M., & Seibert, H. C. (1941). Size and shape of head hair from six racial groups. *Journal of Heredity*, *32*(9), 315-318.

Strauss, M. A. T. (1983). Forensic characterization of human hair. *Microscope*, *31*(1), 15-29.

Takahashi, T., Mamada, A., Kizawa, K., & Suzuki, R. (2015). Age-dependent damage of hair cuticle: contribution of S100A3 protein and its citrullination. *Journal of cosmetic dermatology*.

Tridico, S. R., Koch, S., Michaud, A., Thomson, G., Kirkbride, K. P., & Bunce, M. (2014). Interpreting biological degradative processes acting on mammalian hair in the living and the dead: which ones are taphonomic?. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1796), 20141755.

Trotter, M. (1930). The form, size, and color of head hair in American whites. *American Journal of Physical Anthropology*, *14*(3), 433-445.

Trotter, M. (1936). The hair of the Arabs of central Iraq. *American Journal of Physical Anthropology*, *21*(3), 423-428.

Trotter, M. (1938). A review of the classifications of hair. *American Journal of Physical Anthropology*, *24*(1), 105-126.

Trotter, M. (1939). Classifications of hair color. *American Journal of Physical Anthropology*, *25*(2), 237-260.

Trotter, M., & Dawson, H. L. (1934). The hair of French Canadians. *American Journal of Physical Anthropology*, *18*(3), 443-456.

Trotter, M., & Duggins, O. H. (1948). Age changes in head hair from birth to maturity. I. Index and size of hair of children. *American journal of physical anthropology*, *6*(4), 489-506.

Trotter, M., & Duggins, O. H. (1950). Age changes in head hair from birth to maturity. III. Cuticular scale counts of hair of children. *American journal of physical anthropology*, *8*(4), 467-484.

Trotter, M., Duggins, O. H., & Setzler, F. M. (1956). Hair of Australian aborigines (Arnhem land). *American journal of physical anthropology*, *14*(4), 649-659.

Thibaut, S., Barbarat, P., Leroy, F., & Bernard, B. A. (2007). Human hair keratin network and curvature. *International journal of dermatology*, *46*(s1), 7-10.

Thibaut, S., & Bernard, B. A. (2005). The biology of hair shape. *international Journal of Dermatology*, *44*(s1), 2-3.

Turner, W. (1914). The Aborigines of Tasmania: Part 3: The hair of the head. *Transactions of the Royal Society of Edinburgh*, *50*, 309-347.

Vernall, D. G. (1961). A study of the size and shape of cross sections of hair from four races of men. *American journal of physical anthropology*, *19*(4), 345-350.

From: [redacted]
To: [redacted]
Cc: [redacted]
Subject: Follow-up information for "Forensic Science in the Criminal Courts: Ensuring Scientific Validity Of Feature-Comparison Methods"
Date: Tuesday, December 13, 2016 3:14:26 PM

Members of PCAST,

On behalf of the authors of one of the forensic DNA probabilistic modeling mixture analysis papers cited in your document (Greenspoon SA, Schiermeier-Wood L and Jenkins B. Establishing the Limits of TrueAllele® Casework: A Validation Study. J Forensic Sci. 2015;60(5):1263-1276), we would like to clarify some points addressed by your committee.

1. The paper published by our group at the Virginia Department of Forensic Science (VDFS) was a report of studies performed by us and not directed by or affiliated with the manufacturer of the technology (Cybergenetics). It was an independent assessment of the probabilistic modeling program designed for the purposes of not only validating the system for use on forensic casework, but also to demarcate the limitations.
2. We tested and successfully deconvoluted (separated out the contributor genotypes using TrueAllele® Casework) mixture profiles consisting of two, three and four people. TrueAllele® Casework (TA) successfully and *reproducibly* detected the most minor contributor to three person mixtures where the minor contributor was as low as 8%. For the four person mixtures, TA successfully and *reproducibly* detected the most minor contributor to four person mixtures where the minor contributor was as low as 15%.
3. Perhaps more importantly however, is the fact that when compared to 100 synthetic non-contributor PowerPlex® 16 STR profiles, all of the deconvoluted mixtures described above produced negative log(LR) values except for one sample which was not reproducible.

Thank you for the opportunity to clarify information from the above mentioned article.

Regards,

Linda C. Jackson | Director
Virginia Department of Forensic Science
700 N. 5th Street
Richmond, VA 23219
[redacted]

Note: Correspondence referencing a specific case may be retained and subject to disclosure as part of the case file.

An Appeal for Reconsideration of 1 – 3 studies by the authors of the PCAST Report

Key:

TP = true positive conclusion
FP = false positive conclusion
TN = true negative conclusion
FPR = false positive rate
P = total number of positive or same-source comparisons
N = total number of negative or different-source comparisons

The authors of this report insisted that "set-based" black-box test designs as usually used for firearm-toolmark validity testing were "not appropriate for assessing scientific validity and estimating the reliability because they employed artificial designs that differ in important ways from the problems faced in casework." (p. 106.) I concede that generally speaking the authors are correct insofar as these designs can indeed involve dependencies among the comparisons of specimens. No study is perfect in every minute detail; all studies simply vary in their degree of imperfection. But some of these studies in fact *do not* differ in important ways from actual casework.

In 2003, Doug Murphy and I succeeded in having published in the *AFTE Journal* a cartridge case, black-box study that used a "within set" design (see copy of article appended). With this design we sought to replicate actual casework and to prevent the study from becoming too large and cumbersome. Still, this necessarily means there were possible dependent conclusions reached by the test examiners, thus making confidence interval computations more difficult. That is, if A = B and B = C, then A = C by logical association, and thus an actual microscopic comparison between A and C may have been omitted though the result still recorded. If so, the A:C result is thus considered dependent. In all other respects the study was carefully and appropriately designed. However, and of critical importance, there were no FP or false-negative errors; additionally, the Sensitivity was calculated as 100% (TPs/P's) and the Specificity as 40.7% (TNs/N's).

The PCAST report made no mention of this study, presumably for the reason, principally, of its within-set design. But to dismiss or ignore the study for this reason was too hasty in our view, as demonstrated below. Reasonable and conservative estimates can be made for the total number of independent and dependent conclusions, and confidence-intervals then computed via an online calculator, as suggested in the PCAST report. Moreover, this 2003 study satisfies most of the committee's other criteria for "foundational validity," as the latter were listed in the report:

- The study involved only eight test examiners, but it was large enough such that the results were not meaningless; and though the total comparison count—or sample size—is not equivalent to the test-examiner population, a calculated confidence interval does account for sample size. Then too, studies of similar design, similar content, and with the same zero-error results, in principle, should be capable of having data aggregated for a much larger examiner population and sample size. More on this shortly.

- The study was double blind. Not totally blind, which is very difficult to obtain; but just as in clinical drug testing, and as with the cited Ames cartridge-case study, neither the proctors nor the test subjects knew any specific answers or were capable of learning the answers without examining and de-coding a Master List held under lock and key.

- The study design was set in advance and not modified at any time.

- The study was overseen by myself and Doug Murphy. Both we and the test examiners were employed in the Firearms-Toolmarks Unit of the FBI laboratory, so it cannot be said that the test was completely independent. However, (1) the FBI Latents study cited by PCAST, and judged favorably by them, also was not completely independent and yet was deemed by them as an entirely suitable foundational study, and (2) nevertheless, as one can observe from reading our article, every possible precaution was taken in the study's design to prevent bias or leakage of possible answers. Even the Forensic Science Research Unit was consulted on the design.

- The answer sheets and distribution Master List are technically still available for examination by other serious, and interested, parties.

Without *immediate* access to the answer sheets, and thus in order to estimate the actual minimum number of independent comparisons in this study, I very recently undertook a simple, empirical sampling simulation involving three hypothetical test examiners performing the 2003 test. First, consider the comparisons of positive, or same-source specimens. One "examiner" took what can be considered as a typical test; a second took a test heavily skewed toward dependent comparisons; and a third took a test heavily skewed toward independent comparisons (see appended tables that are very similar to the actual answer sheets provided to the test examiners). The combined proportion of independent comparisons for the three was ~53% (17/32). That is, for all the positive comparisons possible, at a minimum 53% must have been original, "uncontaminated" comparisons. Relevant to these appended tables, note also that for our study the eight examiners were each exposed to an average of ~ nine positive comparisons out of 45 possible comparisons (70 P's total/eight examiners = 8.75 per examiner).

This result goes to Sensitivity. Instead of the calculation TPs/P = 70/70 = 100%, we now have 35/35 = 100%, where we're estimating that half, or 35, of the positive comparisons were independent.

But the real issue is FPR (FPs/N) and the denominator of this ratio, such that we can compute a confidence interval. We know the numerator is zero from this study (no FPs), but how many independent, negative (different-source) comparisons were there? I won't belabor the point with more paper, but when a second analysis is undertaken with the same appended tables for the simulations, it turns out there were 103 total negative comparisons, of which 14 were dependent (assuming no inconclusives), for a percentage of ~14%. But let's be conservative and estimate the proportion of *independent* negative comparisons as not 86% but 70%. For our study, there were 290 different-source comparisons. Seventy percent of this figure would yield 203 independent negative comparisons where there was an opportunity for false-positive conclusions. With zero false-positive errors, what then would be the Clopper-Pearson 95% confidence interval figures? Using the online calculator suggested by the PCAST authors, the result is a 95% confidence interval of from 0.0% to 1.8% (see appended copy of the online results). This result is clearly in the same neighborhood with the 95% confidence interval from the Ames study, 0.6% to 1.5%—when using the standard definition of false-positive rate (FPR). *But the main point here attaches not to the results themselves, but rather to the fact that such a calculation is possible and valid—that accuracy from the test is indeed capable of being conservatively estimated.*

The foregoing analysis invoked the conventional definition for FPR, derived from the four-cell, binary classification system (yes or no). The PCAST authors prefer to use an *ad hoc* definition of FPR, wherein the number of FPs is divided not by the total number of negatives (N) but rather by the number of actual negative *conclusions*—TNs + FNs—in what's a six-cell, trinary classification system (yes, no, maybe). I'm of the opinion that this ad hoc definition is debatable at best,

but nevertheless, using this method, the Clopper-Pearson 95% confidence interval for our study spans the range from 0.0% to 4.3%, whereas the Ames study reports in at 1.0% to 2.3%.

One assertion/criticism the PCAST authors doubtless would make regarding our 2003 study is that a within-set design is not a black-box design (see their Note 335), and only black-box designs go toward validity. On p. 48 they define a black-box study as "an empirical study that assesses a subjective method by having examiners analyze samples and render opinions about the origin or similarity of samples." But to the contrary, that's *exactly* what our and many other studies did. True, the within-set design may involve comparison dependencies, but it's exactly the format examiners are often presented with in actual casework (Here are six cartridge cases. What can you tell us?)

In this same footnote, the PCAST authors go on to criticize for the purpose of validity testing the inclusion of between-class comparisons that existed in a separate study. Our study also involved a distinct minority of between-class comparisons (~ 35 such comparisons, out of the total of 290 N's, or 25 out of 203 if dependencies are removed). If the 25 were subtracted from the 203, the resulting 95% C-P confidence interval would merely be shifted to 0.0% to 2.1%—hardly much of a change. But whether such slicing and dicing is utterly necessary seems open to some debate. It's true that normally only TP conclusions from same-class guns and specimens would ever find their way into court. However, the PCAST authors insist in Note 335 that "the central question ... is whether examiners can associate spent ammunition with a *particular* gun, not simply with a particular *make* of gun." But examiners are commonly asked to determine if there's a link between a bullet or cartridge case fired in a gun with class characteristics different from those of a submitted gun, and the submitted gun. The question to be answered is the same: Was this cartridge case fired in this particular gun? And because class characteristics imparted to cartridge cases from different makes/models of guns are often similar, comparison microscopy is very often conducted in these cases. So for all these circumstances combined, a determination must still be made in answering the same central question, regardless of whether comparison microscopy is conducted. Why should this reality be ignored when tabulating data and calculating accuracy? Our study—and others—most closely mimics real world experience, in which all these categories of laboratory comparisons occur. To draw inferences from studies involving only same-class guns is unrepresentative of actual casework and would skew an *aggregate* error rate.

Even so, I can't completely dispute their logic. If it's possible to more finely tune the research to better reflect our specific environment, in which for the most part only positive results from same-class specimens are heard in court, then some research surely should be designed and conducted on this basis. After all, the most crisp and relevant courtroom question here is, Out of all the different-source, same-class, comparisons conducted by laboratories such as yours, what percentage results in a FP error? That an ID conclusion was effected between same-class specimens is a known piece of information (an ID must involve same-class), and therefore should be used.

But I can also cogently argue that a different design isn't invalidated simply because the design cannot be used to perfectly answer someone's particular question, or the most relevant question. To use an imperfect analogy, the Framingham cardiovascular-disease-risk calculator commonly used by cardiologists accounts for only a few variables such as blood pressure and age when yielding a probability that one will have a heart attack or stroke within ten years. If more was known for a particular patient—for example, that his close relatives all suffer from diabetes, then this information clearly could be used to adjust upward the Framingham output. In theory one could do only research that included diabetes as a variable, and perhaps this would be to the good. But this wouldn't "foundationally" invalidate the existing Framingham database and research, which doesn't fully and completely answer any particular patient's exact question that takes in his/her individual health information.

For us, the global (Framingham-type) question is, What is the FPR flowing out of forensic firearms exams across the world, or at least across the United States?, and the global set of comparisons would include different-class comparisons. Our study and others help to answer this broader question, even if not fully and precisely tailored to the

courtroom environment. In any event, I'm probably making a mountain out of a molehill; as already observed, accounting for PCAST's objection hardly changes the confidence interval of our results.

In sum, my co-author and I would argue that our 2003 Glock cartridge-case study should be seriously reconsidered by the PCAST authors as one that directly and logically supports foundational validity, when a fair appraisal is undertaken with reasonable and conservative estimates for the number of independent conclusions. No, it's not quite as sound in design or as comprehensive as the Ames study; it's not perfect; but it's pretty good and should not be completely discounted and thereby banished to the "Island of Misfit Toys."

Moreover, there were at least four other studies undertaken within the FBI's Firearms-Toolmarks Unit that used our study as a design template. Thus they also suffered from the same within-set dependency problem. Two of these were bullet studies cited by PCAST (their footnotes 319 and 320), a third involved screwdrivers, and a fourth tested examiners' ability to correctly associate fractured surfaces. I'm not as familiar with the details of these studies and therefore haven't commented on them, but I do know there were zero FP errors with the two cited bullet studies. Given this fact, and using the same 70% conservative estimate for the proportion of independent, same-source comparisons, we can combine the number of different-source comparisons from all three studies and arrive at an adjusted, aggregate total of 784. When one computes the Clopper-Pearson confidence interval using the standard FPR definition, the result is 0.00% to 0.47%. These are impressive numbers, and the computation makes use of a significantly larger number of total comparisons.

Steve Bunch
12/12/2016

A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases

By: Stephen G. Bunch and Douglas P. Murphy, FBI Laboratory, Washington DC

Key Words: breechface, cartridge cases, consecutive manufacture, Glock, individuality, validity, validity study

ABSTRACT

Very little comprehensive research has been conducted to date on the scientific validity of traditional forensic cartridge case comparisons. This study is an initial attempt to remedy this scarcity, and was conducted under the assumption that the firearms-toolmarks examiner is an integral part of a scientific process leading to a reported conclusion. Thus, conclusions were drawn by a group of "control" examiners, in this case those in the Firearms-Toolmarks Unit of the FBI Laboratory. The results herein confirm the scientific validity of traditional comparisons.

INTRODUCTION

Our purpose in this study was to rigorously test two propositions: 1) that marks imparted to cartridge cases fired from different guns rarely if ever display sufficient agreement to lead a qualified firearms examiner to conclude the specimens were fired from the same gun, and 2) that marks imparted to cartridge cases from the same gun will rarely if ever lead a qualified firearms examiner to conclude the specimens were fired from different guns. These are the strongest claims commonly made by examiners and are readily tested.

This study does not attempt to assess the relative strengths and weaknesses of the various stages of the firearms identification process, only the final result. As such, the cause of any incorrect response cannot be identified. However, correct responses lend validity to every step in the examination process.

Note that inconclusive results are not addressed here. They do not directly bear on the strong claims, and can vary widely depending on the specimens being examined. While inconclusive results are relevant to training, competency, and quality assurance issues in general, they are not truly relevant to the scientific validity of bullet and cartridge case comparisons, for the simple reason that firearms examiners make no firm, testable claims about them.

PRINCIPLES OF THE TEST DESIGN

This study is designed along familiar proficiency test lines, though there are critical differences that allow this to be more properly termed a validity test, not a quality assurance-type proficiency test:

→ First, test participation and return was mandatory for all qualified firearms examiners in the FBI Laboratory's Firearms-Toolmarks Unit, with the exception of the test administrators. In this way no self-selection bias and "survivorship" bias could be introduced, the latter by an examiner deciding mid-way through the test that he wouldn't turn it in owing to fear of a possible error.

→ The design eliminated the possibility that test examiners could somehow uncover the correct answers. Conversations between examiners would elicit absolutely no useful information. More on this shortly.

→ If any errors were committed, they could not possibly be traced to an individual examiner. Examiners may have a tendency when being tested to become more conservative to reduce the possibility of committing an error. This test's anonymity encouraged examiners to treat the test as an actual case, and refrain from "playing it safe."

→ The test was blind insofar as the test takers (test examiners or "control" examiners) were ignorant of the correct answers. It was double blind insofar as the test administrators (Bunch and Murphy) had no knowledge of which examiner received which test specimens (the anonymity feature again). Further, to learn of specific answers associated with specific test packets would have required the administrators to consult a Master List that was under lock and key. By no means could the administrators have signaled answers to the test takers, either overtly or with subtlety.

Even so, this test was open and could not be considered totally blind. The test examiners knew they were participating in a validity study. Indeed, complete concealment is virtually impossible to achieve. A background "legend" must be created for each evidence test packet, but a curious examiner who asks a few questions will invariably have his suspicions aroused that a submission of evidence could constitute part of a test.

→ Cartridge cases that had been marked by 10 consecutively manufactured Glock pistol breechfaces were used in this study. In theory this made the test

somewhat more severe than would have been the case for actual forensic comparisons. On the other hand, marks imparted to cartridge case primers from Glock breechfaces are, under normal circumstances, readily identifiable. It also should be noted that the Glock pistol strikers were not consecutively manufactured.

→ Answer and instruction sheets provided to the test examiners were to be as simple as possible in order to minimize quality assurance errors, as opposed to scientific errors. Each examiner would simply write the appropriate 2-letter set in the appropriate space on the answer sheet: ID for identification, EL for elimination, and NC for no conclusion (inconclusive). A sample answer sheet is included at the end of this article.

ACQUISITION OF GLOCK PISTOLS

Two firearms examiners from the FBI Laboratory, including one of the test administrators, traveled to Glock's manufacturing facilities in Ferlach and Deutsch-Wagram, Austria in October of 2000 to personally observe the manufacturing process. This served two purposes. First, the consecutive manufacture of the firearm breechfaces could be verified. Second, the types of manufacturing processes could be observed and noted. The most important manufacturing steps we observed were the following:

1) Ten unfinished slides were selected at random.
2) The final breechface dimensions were created using a Computer Numerically Controlled (CNC) single-edged cutter. This cutter makes approximately 60 passes until the correct dimension is achieved and is sharpened in-house approximately every 1000 slides.
3) The extractor grooves were cut with a CNC rotating broach, which appeared to leave some toolmarks on the breechface.
4) The firing pin apertures were punched out from behind the breechface with a tool bearing a ground surface.
5) A slightly wedge-shaped tool the size of a firing pin was used to smooth the sharp edges around the firing pin apertures (CNC also).
6) A light hand-filing operation was performed to remove any remaining burrs and remove the slight bulge around the firing pin apertures.
7) The breechface is protected or unaffected during all of the remaining operations.
8) The firing pins are produced exclusively by grinding operations with no final polish.

PROCEDURES AND TEST DESIGN DETAIL

After initial break-in and cleaning, each of the ten 9mm Luger Glock pistols was fired ten times and the cartridge cases collected in marked "gun bags," one bag per gun. The same occurred for a Beretta model 92F and a SigSauer model P226, both 9mm Luger pistols obtained from the Laboratory's collection, so that the total specimen count was 120. (These latter pistols were included in order to allow for elimination conclusions.) Two randomly selected specimens from each Glock gun bag were then examined under a comparison microscope to ensure that at least some displayed useful, detailed marks on the primers.

Now the cartridge cases in each gun bag were scribed and sticker-labeled with a 3-digit number that was randomly generated using mathematical software. Of course we ensured that no duplicate, triplicate, etc., set of numbers was used. Next we obtained examination packets, numbered 1 – 8, each to be anonymously distributed to a test examiner, and to contain the cartridge cases for examination.

In order to preclude the usefulness of possible conversations between examiners and cover the full range of possible conclusions in this test, the combinations of cartridge cases present in each examiner package was controlled to a certain extent. Eight test examiners participated in the study, so into one of the exam packets were placed ten cartridge cases from a single Glock gun bag. Into another exam packet were placed one cartridge case from each of the nine remaining Glock gun bags, plus one cartridge case from one of the non-Glock bags. Into the remaining six exam packets were placed either zero, one, or two cartridge cases from the non-Glock pistols. (All the while, what went where was recorded on a Master List under the heading, for example, "Exam Packet 5.")

Next, the remaining contents of all the Glock gun bags were blended into a single cartridge case bin. From this bin were selected at random the specimens that were necessary to fill the remaining six exam packets to a total of ten cartridge cases. Once the exam packets were complete, eight answer sheets were prepared, one for each test examiner. Each sheet listed the 3-digit specimen identifying numbers, in grid fashion, indicating the comparisons to be conducted. For every empty box on the sheet, the test examiner would write the letters representing one of the three possible conclusions.

This procedure ensured that no pattern existed between the various exam packets. Conversations between test examiners could not possibly result in a test examiner obtaining any useful information.

Once all the materials were in good order, eight complete test packets were prepared, each containing three items:

the exam packet, the answer sheet specific to that exam packet, and an instruction sheet that set forth the three possible conclusions and their meaning, as well as an admonition to take care to avoid mixing up cartridge case identifying numbers when marking the answer sheet.

Before the distribution of the test packets, the test examiners were briefed on the test's anonymity features, instructed to treat the exam specimens as evidence in a normal case, and given an explanation of the answer sheet. Finally, the test packets were set aside in an "outgoing" box in a closed room, and the test examiners instructed to return them to the "incoming" box when finished. They also were to ensure that neither of the authors saw them going in or out of the closed room (silly, but necessary for completely untainted results).

RESULTS AND DISCUSSION

The total number of comparisons conducted by the test examiners was 360, with 42 of these between cartridge cases fired in consecutively manufactured pistols. There were no mis-identification or mis-elimination errors committed in this study (false positives and false negatives, respectively).

For data analysis purposes, a forensic comparison examination can be considered analogous to a clinical test such as a blood test. The overall quality of these tests is often measured by considering four quantities: the false positive and false negative error rates; a quantity termed *sensitivity*, which is the number of positives actually obtained from a test divided by the number of true positives; and a quantity termed *specificity*, which is the number of negatives actually obtained from a test divided by the number of true negatives.

As observed, the false positive and false negative error rates realized from this study were zero. Additionally, there were 70 true positives, which exactly matched the number of identification conclusions offered—thus the sensitivity figure for this forensic examination, for this study, was 100% (70/70 = 1). There were 290 true negatives, while elimination conclusions offered totaled 118—thus the specificity figure for this forensic examination, for this study, was 40.7% (118/290 = .407).

While false positives and negatives are the most important measures of quality for forensic comparison examinations, sensitivity and specificity are also indicators of a test's quality and should be given due consideration. Firearms and toolmarks comparison examinations, however, do present a complication, given the necessary and proper existence of inconclusive results. Depending on the design of a particular validity study, inconclusive results could be expected to range from a small fraction of total results to a large fraction. (For example, consider the likely results of this study if it had been conducted with *bullets* fired from Glock pistols.) Unlike many clinical tests, there is no inherent tendency for sensitivity and specificity figures to cluster around a "true" figure. It depends on the circumstances, and this is especially true for specificity.

Despite this, the overall results from this study were excellent. Scientifically speaking, the validity of forensic cartridge case comparison examinations was strongly supported. The two propositions directly tested by this study were confirmed. The 100% sensitivity result was noteworthy, and the 40.7% specificity result simply reflects the nature of these examinations and may take on lesser importance than earlier suggested when considered in the caseworking context. That is to say, in actual casework, as opposed to validity testing, the "operational" specificity figure could well be higher, for quite often a submitted cartridge case and gun display incompatible class characteristics if in fact the cartridge case had not been fired in the submitted firearm.

A final and important point to keep in mind is that studies such as this assess the overall scientific validity and quality of examinations (or tests) such as forensic cartridge case comparisons. The results, however, should not be used to arbitrarily estimate error rates in actual laboratory work. Laboratory practice often involves additional quality assurance measures such as the confirmation of identifications and technical peer review. Moreover, the probability of an error in any particular case depends on numerous factors, and is affected by the so-called base rate, or the prior odds of a cartridge case having been fired in a submitted firearm.

FINAL REMARKS

The authors would like to thank Gaston Glock and Robert Glock of the Glock Corporation, as well as Werner Ruppnig and the engineering staff at Glock's Ferlach facility. Their tremendous hospitality and willingness to assist us in any possible way was extremely gratifying and contributed greatly to the success of this project.

Sample test-examiner #1: Typical test. Four same-gun groups. True Positives (TPs) comprise A=E=J, B=C, D=G, F=I=H.
Assume sequence of comparison conclusions runs along the columns, left to right. Thus test-examiner reached following conclusions. In table, dependent TP conclusions are **bolded**, <u>underscored</u>, and in larger font:

A = E

A = J, and therefore E = J

C = B

G = D

H = F

I = F, and therefore H = I

Six of the eight TP conclusions were independent.

A	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
B		XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
C		TP	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
D				XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
E	TP				XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
F						XXXX	XXXX	XXXX	XXXX	XXXX
G				TP			XXXX	XXXX	XXXX	XXXX
H						TP		XXXX	XXXX	XXXX
I						TP		**<u>TP</u>**	XXXX	XXXX
J	TP				**<u>TP</u>**					XXXX
	A	B	C	D	E	F	G	H	I	J

Sample test-examiner #2: Extreme dependency test. One same-gun group. TPs comprise D=E=F=G=H=I.

Five of the fifteen TP conclusions were independent.

A	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
B		XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
C			XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
D				XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
E				TP	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
F				TP	**TP**	XXXX	XXXX	XXXX	XXXX	XXXX
G				TP	**TP**	**TP**	XXXX	XXXX	XXXX	XXXX
H				TP	**TP**	**TP**	**TP**	XXXX	XXXX	XXXX
I				TP	**TP**	**TP**	**TP**	**TP**	XXXX	XXXX
J										XXXX
	A	B	C	D	E	F	G	H	I	J

Sample test-examiner #3: Extreme independency test. Four same-gun groups. TPs comprise A=E=F=J, B=I, C=H, D=G.

Six of the nine TP conclusions were independent.

A	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
B		XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
C			XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
D				XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
E	TP				XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
F	TP				**TP**	XXXX	XXXX	XXXX	XXXX	XXXX
G				TP			XXXX	XXXX	XXXX	XXXX
H			TP					XXXX	XXXX	XXXX
I		TP							XXXX	XXXX
J	TP				**TP**	**TP**				XXXX
	A	B	C	D	E	F	G	H	I	J

ausvet

EpiTools epidemiological calculators

Calculate confidence limits for a sample proportion

Input Values

Sample size : 203

Number positive : 0

Confidence level: 0.95

Confidence Interval method: Clopper-Pearson exact

This utility calculates confidence limits for a population proportion for a specified level of confidence.

Inputs are the sample size and number of positive results, the desired level of confidence in the estimate and the number of decimal places required in the answer.

The program outputs the estimated proportion plus upper and lower limits of the specified confidence interval, using 5 alternative calculation methods decribed and discussed in Brown, LD, Cat, TT and DasGupta, A (2001). Interval Estimation for a proportion. *Statistical Science* **16**:101-133:

1. Asymptotic (Wald) method based on a normal approximation,
2. Binomial (Clopper-Pearson) 'exact' method based on the beta distribution,
3. 'Wilson' Score interval,
4. 'Agresti-Coull' (adjusted Wald) interval and
5. 'Jeffreys' interval.

The Wald Interval often has inadequate coverage, particularly for small n and values of p close to 0 or 1. Conversely, the Clopper-Pearson Exact method is very conservative and tends to produce wider intervals than necessary. Brown et al. recommends the Wilson or Jeffreys methods for

Decimal places in answer : 3

small n and Agresti-Coull, Wilson, or Jeffreys, for larger n as providing more reliable coverage than the alternatives. Also note that the point estimate for the Agresti-Coull method is slightly larger than for other methods because of the way this interval is calculated.

Submit

Confidence limits for a proportion

Analysed: Mon Dec 12, 2016 @ 22:24

Inputs

Sample size	203
Number positive	0
Confidence level	0.95
CI method	Clopper-Pearson exact

Results

	Number positive	Sample size	Proportion/Prevalence	Lower 95% CL	Upper 95% CL
Clopper-Pearson exact	0	203	0.000	0.000	0.018

CI plot

ausvet

EpiTools epidemiological calculators

Calculate confidence limits for a sample proportion

Input Values

Sample size : 83 → = 0.7(118)

Number positive : 0

Confidence level: 0.95

Confidence interval method: Clopper-Pearson exact

This utility calculates confidence limits for a population proportion for a specified level of confidence.

Inputs are the sample size and number of positive results, the desired level of confidence in the estimate and the number of decimal places required in the answer.

The program outputs the estimated proportion plus upper and lower limits of the specified confidence interval, using 5 alternative calculation methods decribed and discussed in Brown, LD, Cat, TT and DasGupta, A (2001). Interval Estimation for a proportion. *Statistical Science* **16**:101-133:

1. Asymptotic (Wald) method based on a normal approximation,
2. Binomial (Clopper-Pearson) 'exact' method based on the beta distribution,
3. 'Wilson' Score interval,
4. 'Agresti-Coull' (adjusted Wald) interval and
5. 'Jeffreys' interval.

The Wald interval often has inadequate coverage, particularly for small n and values of p close to 0 or 1. Conversely, the Clopper-Pearson Exact method is very conservative and tends to produce wider intervals than necessary. Brown et al. recommends the Wilson or Jeffreys methods for

Decimal places in answer : 3

small n and Agresti-Coull, Wilson, or Jeffreys, for larger n as providing more reliable coverage than the alternatives. Also note that the point estimate for the Agresti-Coull method is slightly larger than for other methods because of the way this interval is calculated.

Submit

Confidence limits for a proportion

Analysed: Mon Dec 12, 2016 @ 22:32

Inputs

Sample size	83
Number positive	0
Confidence level	0.95
CI method	Clopper-Pearson exact

Results

	Number positive	Sample size	Proportion/Prevalence	Lower 95% CL	Upper 95% CL
Clopper-Pearson exact	0	83	0.000	0.000	0.043

CI plot

From: [redacted]
To: [redacted]
Subject: Forensic Cartridge Case Study for PCAST
Date: Tuesday, December 13, 2016 4:17:51 PM
Attachments: Response to PCAST solicitation for validity studies.pdf

---

Dr. Lander:

Originally the Virginia Dept. of Forensic Science (DFS) was thinking of putting together a response to PCAST's "solicitation," and as part of that we in the Dept. were given the opportunity to reply/comment, at which point the Dept. would collate the replies and forward them to you. However, I perhaps misunderstood the Dept's intent. The bottom line is, the DFS is not going to be responding as an organization, as I just learned today. Instead, the Dept. encouraged me to send the document directly to you and the PCAST panel.

So in this vein, the attached scanned document is what I had sent yesterday to the managers here in the Dept... So they, perhaps more than the PCAST panel, were the audience. Please inform any reviewers of same, because I have no time to do any revisions before the deadline for "submissions." Also, as required, I don't believe the PCAST panel reviewed this study; at least there are no such citations in your final report.

The document attached was prepared by me without a lot of time, so in the document itself I didn't drill down into the details about estimating the number of different-source, independent and dependent conclusions, etc. But I have kept my notes just in case.

I'm hoping that the committee calls on the ghost of Enrico Fermi, haha, when thinking of this. Of course, as you know, he was the originator of "Fermi Estimation," and therefore I hope the fact that I'm invoking what I think are good estimations isn't immediately a show stopper (even though I'm not exactly doing Fermi estimations). Especially since my estimations are conservative. They're especially conservative also in a way I didn't mention in the document—that as a matter of the actual test taking, I'm very confident the test examiners did not use, or rarely used, logic to group specimens and thus effect dependent conclusions. Still, it's not possible to estimate this, and so I didn't include it as a factor.

Finally, I must issue a disclaimer. Clearly this document represents my views and perhaps those of my co-author, and not necessarily the views of the VA Dept. of Forensic Science.

Thank you for your time and consideration,

Stephen G. Bunch, Ph.D.
Firearms & Toolmarks Section Supervisor
Virginia Department of Forensic Science - Northern Lab
10850 Pyramid Place
Manassas, VA 20110
[redacted]

Stephen.Bunch@dfs.virginia.gov

Note: Correspondence referencing a specific case may be retained and subject to disclosure as part of the case file.

THE AMERICAN CONGRESS OF FORENSIC SCIENCE LABORATORIES

The United States Assembly of Forensic Science Laboratory Professionals

Our Mission
To represent and unite all current and former professionals of United States forensic science laboratories with the purpose of creating and preserving the conditions necessary for the American criminal and civil justice systems to have confidence in the integrity of forensic laboratory services.

The American Congress of Forensic Science Laboratories
c/o The Forensic Foundations Group
901 S. Bridge Street, Number 227
Dewitt, MI 48823
(517) 803-4063
office@forensicfoundations.com

December 14, 2016

Dr. Eric Lander
c/o US Department of Justice
The Office of Legal Policy
950 Pennsylvania Avenue NW
Washington, DC 20530

Dear Dr. Lander:

The American Congress of Forensic Science Laboratories is in receipt of an invitation (which was not dated but was received by email on December 2, 2016) in which our organization was asked to provide you with foundational research in support of the several forensic laboratory sciences criticized by the recent report issued by the President's Council of Advisors on Science and Technology (PCAST).

We appreciate the opportunity to assist you, but we are respectfully declining your invitation. While we respect the Office of the President of the United States and the experts it convenes to examine a wide variety of issues relevant to our fellow citizens, we believe that PCAST is not an entity having sufficient authority or relevance such that it can summon an entire scientific community to justify its own existence in less than two weeks using an unnecessarily strict definition of foundational validity.

In this regard, we encourage PCAST to reassess its criteria for what constitutes foundational validity. Science is a highly accommodating institution, yet PCAST subscribes to a remarkably constrictive and myopic definition of science as if to strategically discredit as many forensic science disciplines as possible.

Moving forward, it is our opinion that PCAST's intent to function in good faith will be indicated by its willingness to keep an open mind and engage a wide variety of professionals working in forensic science organizations not only in the United States but around the world. This means visiting crime laboratories where forensic science is practiced, allowing yourself to be introduced to and educated by the people who perform this work, and listening carefully to the many – sometimes differing – perspectives held by those who've worked and managed in the forensic sciences for many years. It is our understanding that PCAST has already been provided significant volumes of information from a variety of forensic science organizations and experts. We hope you give it the serious consideration it deserves.

If we can assist you in improving your approach to learning about the forensic sciences, we would sincerely like to do so. Please do not hesitate to contact us.

Respectfully Submitted,

Kermit Channell
President

AMERICAN SOCIETY OF CRIME LABORATORY DIRECTORS, INC.

139 A Technology Drive Garner, NC 27529

December 13, 2016

Dr. Lander and distinguished members of PCAST,

The ASCLD Board of Directors appreciates your letter dated 12/1/2016 asking for additional input into PCAST's recent report, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods.* As you are probably aware, ASCLD represents more than 600 of the nation's crime laboratory directors and managers, and we take seriously any opportunity to provide the perspective of crime laboratory administrators and practitioners and inform non-forensic stakeholders into the needs and state of forensic science.

While we truly appreciate your request, we are unable to provide you the requested information as of this date due to the extremely short deadline you assigned. ASCLD notes that PCAST took more than a full year to develop the report with more than 100 revisions. Two weeks is simply not enough time to provide thorough, well-reasoned responses to your questions, particularly during this time of year where laboratory directors are very busy with year-end operational issues.

We hope that in the future PCAST will provide another opportunity for ASCLD's input into the Report with a more realistic response timeline.

Kindest regards,

Jeremy Triplett
ASCLD President

Phone: 919.773.2044 | Fax: 919.861.9930 | *Website: www.ascld.org*

Association of

Firearm and Tool Mark Examiners

December 13, 2016

Dear Dr. Lander,

In late November 2015 the President's Council of Advisors on Science and Technology (PCAST) posted a seven-question survey on the White House website for public input on forensic science in the United States, with any comments due by December 23, 2015. Specifically, PCAST was inquiring about research supporting the foundational validity of a range of pattern-based forensic disciplines.

In response to this invitation, the Association of Firearm and Tool Mark Examiners (AFTE) submitted a document highlighting more than forty papers authored by AFTE members, academics, or physical science researchers and published in the *Journal of Forensic Sciences*, the *AFTE Journal*, *Forensic Science International,* and other publications in the last few years. Several of these studies were federally funded by the US Department of Justice (DOJ), the National Institute of Justice (NIJ), or the US Department of Energy (DOE). The list provided by AFTE only scratched the surface of the research that has been performed to assess the validity and reliability of firearm and toolmark identification over numerous decades, and the document directed PCAST to find additional information neatly packaged in the Association's online Admissibility Resource Kit: https://afte.org/resources/swggun-ark.

The PCAST Report to the President on "Forensic Science in the Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" released in September 2016 declared that nearly every forensic science method discussed failed to meet their scientific criteria. AFTE released a statement in October 2016 addressing its concerns with the report. Most recently, a form letter from PCAST Co-Chair Eric Lander (which was undated but distributed via email on December 2) is once again requesting information about the foundational literature of firearms analysis, to be submitted to PCAST no later than December 14, 2016. It is clear to the AFTE Board of Directors that if the previous materials presented in good faith to PCAST have been disregarded, no additional evidence will alter their collective opinion.

AFTE Board of Directors

From: [redacted]
To: [redacted]
Subject: Re: Invitation to Provide Follow-up Information to PCAST Regarding its Forensics Report
Date: Thursday, December 8, 2016 7:26:34 PM

Dr. Lander,

I received your email inviting additional input on the PCAST report on Forensic Science Friday afternoon, and I noted the requested quick turnaround of eight working days to get any additional information to you.

I encourage you to approach this project cooperatively with experts in the relevant forensic disciplines. Practicing forensic scientists in the OSAC subcommittees (Biological Data Interpretation and Reporting, Footwear and Tire, and Firearm and Tool Mark) should be willing to work with you and your committee in providing relevant data and references in support of foundational validity of their disciplines. I would not expect them to be able to produce an exhaustive list of references in eight days.

I would hope that you would not put unreasonable restrictions on the form of the data that you are willing to accept and consider. There is a wealth of data on DNA mixture interpretation, but if you are only willing to consider published data in scientific journals that deals with three person mixtures with the minor contributor below 20%, then you are probably not going to get much of a response.

Thousands of papers and publications relating to firearms comparison exist. It would be a matter of someone taking the time to go through and determine which ones speak to foundational validity. Again, if the effort is cooperative it has a chance of providing you with information you could find useful.

This is a project that should be undertaken by PCAST with assistance and input by experts in the forensic field, with an open mind and without unreasonably tight deadlines.

Mike Grubb
Crime Laboratory Director
San Diego Co. Sheriff's Dept. Crime Laboratory
[redacted]

Firearms and Toolmark references not cited in the PCAST report

Source: https://afte.org/resources/swggun-ark

Emerging research

Zhang, S. and Chumbley, L.S., "Manipulative Virtual Tools for Tool Mark Characterization", NCJRS Document #241443, Award # 2009-DN-R-119, March 2013

Song, J., et al, "Development of Ballistics Identification- from Image Comparison to Topography Measurement in Surface Metrology", Measurement Science and Technology, Volume 23, Number 054010, March, 2012.

Chu, W., et al, "Selecting Valid Correlation Areas for Automated Bullet Identification System Based on Striation Detection", Journal of Research of the National Institute of Standards and Technology, Volume 116, Number 3, May-June 2011.

Error Rates

Murphy, D., "CTS Error Rates, 1992-2005 Firearms/Toolmarks", Presented at the 41st Association of Firearm and Tool Mark Examiners (AFTE) Training Seminar, Henderson, NV, May 5, 2010.

CTS Results Revisited: A Review and Recalculation of the Peterson and Markham Findings, by: Bunch, Stephen.

Murdock and Grzybowski - Firearm/Toolmark Identification- Meeting the Daubert Challenge, AFTE Journal Winter 1998; 30(1):3-14.

Firearm and Toolmark Identification, Biasotti and Murdock, Chapter 23, MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY, By: David L. Faigman, David H. Kaye, Michael J. Saks & Joseph Sanders.

Peterson J.L., Markham P.N., Crime laboratory proficiency testing results, 1978-1991, I: Identification and classification of physical evidence**.** Journal of Forensic Science. 1995 Nov;40(6):994-1008

Peterson J.L., Markham P.N., Crime laboratory proficiency testing results, 1978-1991, II: Resolving questions of common origin. Journal of Forensic Science. 1995 Nov;40(6):1009-1029.

Firearm and Toolmark Identification – Theoretical

Heard, B. J., "Handbook of Firearms and Ballistics", Wiley & Sons, 1997, pp. 136-141

Howitt D., Tulleners F., "A Calculation of the Theoretical Significance of Matched Bullets", Journal of Forensic Sciences, Volume 53, Number 4, July 2008, Pp.868-875.

May L., "Identification of Knives, Tools and Instruments", Journal of Police Science (no volume or number listed) 1930, pp. 247-248.

Neel M., and Wells M., "A Comprehensive Statistical Analysis of Striated Tool Mark Examinations Part I: Comparing Known Matches and Known Non-Matches", AFTE Journal, Volume 39, (4), Summer 2007, pp. 176-198.

Stone, Rocky, "How Unique are Impressed Marks", AFTE Journal, vol. 35 (4), Fall 2003, pp. 376-383.

Firearm Identification – Bullets

Bachrach, B., "Development of a 3D-Based Automated Firearms Evidence Comparison System", Journal of Forensic Sciences, Vol. 47(6), November 2002, pp. 1253-1264.

Biasotti, A. A., "A Statistical Study of the Individual Characteristics of Fired Bullets", Journal of Forensic Sciences, Vol. 4(1), January 1959, pp. 34-50.

Brown, C., Bryant. W., "Consecutively Rifled Gun Barrels Present in Most Crime Labs", AFTE Journal, Vol. 27, No. 3, July 1995, pp. 254-258.

Bunch, S. G. "Consecutive Matching Striation Criteria: A General Critique", Journal of Forensic Sciences, vol. 45 (5), Sept. 2000, pp. 955-962.

Chu, et al, "Automatic Identification of Bullet Signatures Based on Consecutive Matching Striae (CMS) Criteria", Forensic Science International, 231, 2013, Pp. 137-141.

Fadul, T. G., "An Empirical Study to Evaluate the Repeatability and Uniqueness of Striations/Impressions Imparted on Consecutively Manufactured Glock EBIS Gun Barrels", AFTE Journal, Volume 43, Number 1, Winter 2011, Pp. 37-44.

Freeman, R. A., "Consecutively Rifled Polygon Barrels", AFTE Journal, vol.10 (2), June 1978, pp.40-42.

Hall, E. "Bullet Markings from Consecutively Rifled Shilen DGA Barrels", AFTE Journal, vol. 15(1), Jan., 1983, pp. 33-53.

Intelligent Automation, Incorporated, "A Statistical Validation of the Individuality of Guns Using High Resolution Topographical Images of Bullets", National Institute of Justice Grant #2006-DN-BX-K030, October, 2010

Lomoro, V. J., "Class Characteristics of 32 SWL, FIE Titanic Revolvers", AFTE Journal, vol. 6 (2), 1974, pp. 18-21.

Lutz, M., "Consecutive Revolver Barrels", AFTE Newsletter #9, Aug., 1970, pp.24-28.

Matty, W., "A Comparison of Three Individual Barrels Produced from One Button-Rifled Barrel Blank", AFTE Journal, vol. 17(3), July, 1985, pp. 64-69.

Miller, J., "An Examination of Two Consecutively Rifled Barrels and a Review of the Literature", AFTE Journal, vol. 32 (3), Summer, 2000, pp.259-270.

Miller, J., "Criteria for Identification of Toolmarks, Part II: Single Land Impression Comparisons", AFTE Journal, vol. 32 (2), Spring, 2000, pp. 116-131.

Miller, J., "An Examination of the Application of the Conservative Criteria for Identification of Striated Toolmarks Using Bullets Fired from Ten Consecutively Rifled Barrels", AFTE Journal, vol. 33 (2), Spring, 2001, pp. 125-132.

Miller, J., McLean M., "Criteria for Identification of Toolmarks", AFTE Journal, vol. 30 (1), 1998, pp.15-61.

Murdock, J. E., "A General Discussion of Gun Barrel Individuality and an Empirical Assessment of the Individuality of Consecutively Button Rifled .22 Caliber Rifle Barrels", AFTE Journal, vol. 13 (3), 1981, pp. 84-95.

Skolrood, R. W., "Comparison of Bullets fired from Consecutively Rifled Cooey .22 calibre Barrels", Canadian Society of Forensic Science, vol. 8(2), 1975, pp. 49-52.

Smith, E., "Cartridge Case and Bullet Comparison Validation Study with Firearms Submitted in Casework", AFTE Journal, vol. 37 (2), Spring 2005, pp. 130-135.

Tulleners, F., Guisto M., "Striae Reproducibility on Sectional Cuts of One Thompson Contender Barrel", AFTE Journal, vol. 30(1), 1998, pp. 62-81.

Tulleners, F., Hamiel J., "Sub Class Characteristics of Sequentially Rifled .38 Special S&W Revolver Barrels", AFTE Journal, vol. 31 (2), 1999, pp. 117-222.

Firearm Identification – Cartridge Cases

Chu, Tong and Song, "Validation Tests for the Congruent Matching Cells (CMC) Method Using Cartridge Cases Fired with Consecutively Manufactured Pistol Slides", AFTE Journal, Volume 45, Number 4, Fall 2013, pp. 361-366.

Hamby, J., Norris, S., and Petraco, N., "Evaluation of GLOCK 9 mm Firing Pin Aperture Shear Mark Individuality Based on 1,632 Different Pistols by Traditional Pattern Matching and IBIS Pattern Recognition", Journal of Forensic Science, Volume 61, #1, January 2016, pp. 170-176.

Bunch, S. G., Murphy D., "A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases", AFTE Journal, vol. 35 (2), Spring 2003, pp. 201-203.

Coffman, B. C., " Computer Numerical Control (CNC) Production Tooling and Repeatable Characteristics on Ten Remington Model 870 Production Run Breech Bolts", AFTE Journal, Volume 35, Number 1, Winter 2003, pp. 49-54.

Coody, A. C., "Consecutively Manufactured Ruger P-89 Slides", AFTE Journal, Volume 35, Number 2, Spring 2003, pp. 157-160.

Gouwe J., Hamby J. E., Norris, S., "Comparison of 10,000 Consecutively Fired Cartridge Cases from a Model 22 Glock .40 S&W Caliber Semiautomatic Pistol", AFTE Journal, Volume 40, Number 1, Winter 2008, pp. 57-63.

Grooss, K. D., "The 'Hammer-Murderer'", AFTE Journal, vol. 27 (1), 1995, pp. 27-30.

Hamby J., and Thorpe J., "The Examination, Evaluation and Identification of 9mm Cartridge Cases Fired from 617 Different GLOCK Model 17 & 10 Semiautomatic Pistols", AFTE Journal, Volume 41(4), Fall 2009, Pp. 310-324.

Kennington, R., "Identification of Cartridge Cases Fired in Different Firearms: "Pre-Identified Cartridges"", AFTE Journal, vol. 31(1), 1999, pp. 15-19.

LaPorte, D., "An Empirical Validation Study of Breechface Marks on .380 ACP Caliber Cartridge Cases Fired from Ten Consecutively Finished Hi-Point Model C9 Pistols", AFTE Journal, Volume 43, Number 4, Fall 2011.

Lardizabal, P., "Cartridge Case Study of the HK USP", AFTE Journal, vol. 27 (1), Jan., 1995, pp. 49-51.

Lopez, L., Grew S., "Consecutively Machined Ruger Bolt Faces", AFTE Journal, vol. 32 (1), 2000, pp. 19-24.

Lyons, D. J., "The Identification of Consecutively Manufactured Extractors", AFTE Journal, Volume 41, Number 3, Summer, 2009, pp.246-256.

Matty, W., "Raven .25 Automatic Pistol Breech Face Tool Marks", AFTE Journal, vol. 16 (3), 1984, pp. 57-60.

Matty, W., Johnson T., "A Comparison of Manufacturing Marks on Smith & Wesson Firing Pins", AFTE Journal vol. 16 (3), 1984, pp. 51-56.

Mayland and Tucker, "Validation of Obturation Marks in consecutively Reamed Chambers", AFTE Journal, Volume 44, No. 2, Spring, 2012, pp.167-169.

Petraco D. K., et al, "Application of Machine Learning to Toolmarks: Statistically Based Methods for Impression Pattern Comparisons", NIJ/NCJRS Document #239048, Award #2009-DN-BX-K041, July 2012

Rosati, C., "Examination of Four Consecutively Manufactured Bunter Tools", AFTE Journal, vol. 32 (1), 2000, pp. 49-50.

Saribey, A. Y., Hannam A. G., Tarimci C., "An Investigation into Whether or Not the Class and Individual Characteristics of Five Turkish Manufactured Pistols Change During Extensive Firing", Journal of Forensic Sciences, Volume 54, Number (5), September 2009, Pp.1068-1072.

Thompson, E., "Phoenix Arms (Raven) Breechface Toolmarks", AFTE Journal, vol. 26 (2), 1994, pp. 134-135.

Thompson, E., "False Breechface ID's", AFTE Journal, vol. 28 (2), April, 1996, pp. 95-96.

Thompson, R., Song J., Zheng A., and Yen J. "Cartridge Case Signature Identification Using Topography Measurements and Correlations: Unification of Microscopy and Objective Statistical Methods", National Institute of Standards and Technology, Presented at the 18th European Network of Forensic Science Institutes (ENFSI) Conference, Lisbon, Portugal, October, 2011

Uchiyama, T., "Similarity among Breech Face Marks Fired from Guns with Close Serial Numbers", AFTE Journal, vol. 18 (3), 1986, pp. 15-52.

Toolmark Identification

Spotts, R., Chumbley, L.S., "Objective Analysis of Impressed Chisel Toolmarks", Journal of Forensic Science, Volume 60, #6, November 2015, pp. 1436-1440.

Spotts, R., et al, "Optimization of a Statistical Algorithm for Objective Comparison of Toolmarks", Journal of Forensic Science, Volume 60, #2, March 2015, pp. 303-314.

Petraco, N.D.K., et al, "Estimation of Striation Pattern Identification Error Rates by Algorithmic Methods", AFTE Journal, Volume 45, #3, Summer 2013, Pp. 235-244.

Bachrach B., Jain A., Jung S., Koons R.D., "A Statistical Validation of the Individuality and Repeatability of Striate Tool Marks: Screwdrivers and Tongue and Groove Pliers", Journal of Forensic Sciences, Volume 55, Number 2, March 2010, Pp 348-357.

Bacharach, B., "Statistical Validation on the Individuality of Tool Marks Due to the Effect of Wear, Environment Exposure and Partial Evidence", NIJ/NCJRS Document #227929, August, 2009.

Burd, David Q., Gilmore A. E., "Individual and Class Characteristics of Tools", Journal of Forensic Sciences, vol. 13 (3), July, 1968, pp. 390-396.

Butcher, S., Pugh D., "A Study of Marks made by Bolt Cutters", Journal of the Forensic Science Society, vol. 15 (2), Apr., 1975, pp. 115-126.

Cassidy, F., "Examination of Toolmarks from Sequentially Manufactured Tongue and Groove Pliers", Journal of Forensic Sciences, vol. 25 (4), Oct., 1980, pp. 796-809.

Chumbly, L. S., et al, "Validation of Tool Mark Comparisons Obtained Using a Quantitative, Comparative, Statistical Algorithm", Journal of Forensic Sciences, Volume 55, Number 4, July 2010, Pp. 953-961.

Clow, C. M., "Cartilage Stabbing with Consecutively Manufactured Knives: A Response to Ramirez v. State of Florida", AFTE Journal, vol. 37 (2), Spring, 2005, pp. 86-116.

Eckerman, S. J., "A Study of Consecutively Manufactured Chisels", AFTE Journal, vol. 34 (4), Fall 2002, pp. 379-390.

Ekstrand, et al, "Virtual Tool Mark Generation for Efficient Striation Analysis", Journal of Forensic Sciences, Volume 59, #4, July 2014, Pp. 950-959.

Flynn, E. M., "Toolmark Identification", Journal of Forensic Sciences, vol. 2 (1), Jan., 1957, pp.95-106.

Giroux B. N., "Empirical and Validity Study: Consecutively Manufactured Screwdrivers", AFTE Journal, Volume 41, Number 2, Spring 2009, pp. 153-158.

Hall, J., "Consecutive cuts made by bolt cutters and their effect on identification", AFTE Journal, vol. 24 (3), July, 1992, pp. 260-272.

Hornsby, B., "MCC Bolt Cutters", AFTE Journal, vol. 21 (3), July, 1989, p. 508.

Jordan, T., "Individual Characteristics on Copper Insulated Wire", AFTE Journal, vol. 14 (1), 1982, pp. 53-56.

Lee, S. E., "Examination of Consecutively Manufactured Slotted Screwdrivers", AFTE Journal, vol. 35 (1), Winter, 2003, pp. 66-70.

Miller, J., Beach, G., "Toolmarks: Examining the Possibility of Subclass Characteristics", AFTE Journal, vol. 37 (4), Fall 2005, pp. 296-345.

Miller, J., "Cut Nail Manufacturing and Toolmark Identification", AFTE Journal, vol. 30 (3), Summer 1998, pp. 492-498.

Murdock, J. E., "The Individuality of Tool Marks Produced by Desk Staplers", AFTE Journal, vol. 6 (5), 1974. pp. 23-39.

Petraco D. K., et al, "Application of Machine Learning to Toolmarks: Statistically Based Methods for Impression Pattern Comparisons", NIJ/NCJRS Document #239048, Award #2009-DN-BX-K041, July 2012

Reitz, J., "An Unusual Toolmark Identification Case", AFTE Journal, vol. 7 (3), Dec., 1975, pp. 40-43.

Thompson, E. and Wyant, R., "Knife Identification Project (KIP),", AFTE Journal, vol. 35 (4), Fall 2003, pp. 366 "" 370.

Tuira, Y. J., "Tire Stabbing with Consecutively Manufactured Knives", AFTE Journal, vol. 14 (1), 1982, pp. 50-52.

Van Dijk, T. M., "Steel Marking Stamps: Their Individuality at the Time of Manufacture", Journal of the Forensic Science Society, vol. 25 (4), July/Aug, 1985, pp. 243-253.

Watson, D., "The Identification of Toolmarks produced from consecutively manufactured knife blades in soft plastics", AFTE Journal, vol. 10 (3), September, 1978, pp. 43-45.

Watson, D. J., "The Identification of Consecutively Manufactured Crimping Dies", AFTE Journal, vol. 10, September 1978, pp. 19-21.

Zheng, X.A., et al, "2D and 3D Topography Comparisons of Toolmarks Produced from Consecutively Manufactured Chisels and Punches", AFTE Journal, Vol. 46, No. 2, Spring 2014, Pp. 143-147.

National District Attorneys Association
1400 Crystal Drive, Suite 330, Arlington, VA 22202
703.549.9222/703.836.3195 Fax
www.ndaa.org

December 14, 2016

Office of Science and Technology Policy
Executive Office of the President
President's Council of Advisors on Science and Technology (PCAST)
Eisenhower Executive Office Building
1650 Pennsylvania Avenue
Washington, DC 20504

Reference: PCAST's Criteria for the Foundational Validity of Feature Comparison Methods - Six Problems

Dear Co-Chair Holdren and Co-Chair Lander:

On behalf of the National District Attorneys Association (NDAA), the nation's oldest and largest prosecutor organization, representing 2,500 elected and appointed District Attorneys across the United States, as well as 40,000 assistant district attorneys, I write to you again regarding a report recently released on forensic science disciplines. NDAA received a solicitation for any additional relevant scientific studies the September 2016 Report to the President-Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods ("the Report") might have missed.

In its solicitation for any relevant scientific studies that may have been missed in its Report, PCAST proposes to send forensic scientists and other interested parties on a "fool's errand" to provide it with "appropriately designed, research studies" that contain empirical evidence establishing the foundational validity and an estimation of the accuracy of certain feature comparison methods. PCAST's request is essentially designed to lure critics of its Report into providing studies that will inform certain premises contained within the Report (essential foundational validation criteria) that we, and others, dispute. Once collected, these supplemental studies are sure to be rejected as failing to offer sufficient support for the foundational validity of questioned methods because, according to PCAST's own criteria, these studies will not have been "appropriately designed."

Specifically, PCAST asks that we, and others, "identify any relevant scientific reports that (i) have been published in the scientific literature, (ii) were not mentioned in the PCAST report; and (iii) describe ***appropriately designed,*** research studies that provide empirical evidence establishing the foundational validity and estimating the accuracy of any of the following forensic feature-comparison methods, as they are currently practiced" for the five listed disciplines. PCAST then asks that we, and others, "indicate how the scientific reports establish foundational validity and estimate the accuracy of the relevant method" (PCAST solicitation email). (Emphasis added).

PCAST's self-constructed definition of an "***appropriately designed***" research study can be found on pages 52-53, Box 4, of its Report. On this topic, PCAST asserts the following:

> **BOX 4. Key criteria for validation studies to establish foundational validity**
>
> Scientific validation studies—intended to assess the validity and reliability of a metrological method for a particular forensic feature-comparison application— ***must satisfy a number of criteria.***
>
> (1) The studies ***must involve a sufficiently large number of examiners*** and ***must be based on sufficiently large collections of known and representative samples from relevant populations*** to ***reflect the range of features or combinations of features that will occur in the application***. In particular, the sample collections should be:
>
> (a) ***representative of the quality of evidentiary samples seen in real cases***. (For example, if a method is to be used on distorted, partial, latent fingerprints, one must determine the *random match probability*—that is, the probability that the match occurred by chance—for distorted, partial, latent fingerprints; the random match probability for full scanned fingerprints, or even very high quality latent prints would not be relevant.)
>
> (b) ***chosen from populations relevant to real cases***. For example, for features in biological samples, the false positive rate should be determined for the overall US population and for major ethnic groups, as is done with DNA analysis.
>
> (c) ***large enough to provide appropriate estimates of the error rates***.
>
> (2) The ***empirical studies should be conducted so that neither the examiner nor those with whom the examiner interacts have any information about the correct answer***.
>
> (3) The ***study design and analysis framework should be specified in advance***. In validation studies, it is inappropriate to modify the protocol afterwards based on the results
>
> (4) The ***empirical studies should be conducted or overseen by individuals or organizations that have no stake in the outcome of the studies***.
>
> (5) Data, software and ***results from validation studies should be available to allow other scientists to review the conclusions***.
>
> (6) To ensure that conclusions are reproducible and robust, there should be ***multiple studies by separate groups reaching similar conclusions***.

(PCAST Report, pp. 52-53).

There are a number of problems with these criteria. The first is that they are hopelessly (and may have been purposefully left) vague, allowing PCAST the flexibility to later proclaim that the samples utilized in the studies they were/will be provided were not large enough, representative enough, or relevant enough to satisfy their unquantified directives. As such,

they provide forensic practitioners and interested parties with no *a priori* notice or guidance on PCAST's expectations regarding the critical question of "how much is good enough?"

Examples include requirements concerning, "a *sufficiently large* number of examiners" (criteria #1); "*sufficiently large* collections of *representative* samples" (criteria #1); and sample collections "*large enough* to provide appropriate estimates of the error rates" (criteria (1)(c)). Exactly what is "sufficiently large," "representative," or "large enough" is apparently for PCAST's *post hoc* determination, unrestrained by any limitations that would either curtail its judgment or mediate a disagreement on these questions.

The second problem with these criteria is their purported applicability as a *mandatory* "gateway" through which forensic validation studies must pass to be deemed "appropriately designed." The problem with this assertion is that the clinical sources PCAST uses to construct its analogy between the medical field and forensic science explicitly state that the approaches and designs set forth in those documents are *not* the only ones that are scientifically legitimate. Furthermore, these documents explicitly state that their recommendations are *not mandatory or binding on either their authors or the public.*

PCAST claims that "[s]cientific validation studies—intended to assess the validity and reliability of a metrological method for a particular forensic feature-comparison application—*must satisfy* a number of criteria" (PCAST Report, p. 52) (Emphasis added). In footnote 118, PCAST asserts, "*The analogous situation* in medicine is a clinical trial to test the safety and efficacy of a drug for a particular application." (Emphasis added). In support of this proposition, PCAST cites a small handful of sources. Among these are: (1) "Design Considerations for Pivotal Clinical Investigations for Medical Devices: Guidance for Industry, Clinical Investigators, Institutional Review Boards and Food and Drug Administration Staff," issued November 7, 2013, by the FDA, Center for Devices and Radiological Health, and the Center for Biologic Evaluation and Research; (2) "Adaptive Designs for Medical Device Clinical Studies: Guidance for Industry and Food and Drug Administration Staff," issued July 27, 2016, by the FDA, Center for Devices and Radiological Health, and Center for Biologics Evaluation and Research; and (3) "Guidance for Industry E9 Statistical Principles for Clinical Trials," issued in September 1998, by the FDA, Center for Drug Evaluation and Research, and Center for Biologics Evaluation and Research.

A cursory review of these documents reveals that PCAST's claim that "the FDA *requires*" certain criteria contained therein is not exactly true. The caption of each page of the first two cited documents is emblazoned with the following disclaimer, "**Contains Non-Binding Recommendations.**" Further, pages four and one, respectively, of the first two cited documents contains a box that states, in full:

> This guidance represents the Food and Drug Administration's (FDA's) ***current thinking*** on this topic. ***It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative***

> ***approach, contact the FDA staff responsible for implementing this guidance.*** If you cannot identify the appropriate FDA staff, call the appropriate number listed on the title page of this guidance.

(Emphasis added).

Similarly, the first page of "Statistical Principles for Clinical Trials," contains nearly identical language, to wit:

> This guidance represents the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. ***An alternative approach may be used if such approach satisfies the requirements of the applicable statutes and regulations.***

In addition, the Adaptive Design guidance document states, "The use of the word ***should*** in Agency guidance ***means that something is suggested or recommended, but not required***" (Page 2). Furthermore, the Recommendations for Design Considerations for Medical Devices document states:

> Although the Agency has articulated policies related to design of studies intended to support specific device types, and a general policy of tailoring the evidentiary burden to the regulatory requirement, ***the Agency has not attempted to describe the different clinical study designs that may be appropriate to support a device pre-market submission, or to define how a sponsor should decide which pivotal clinical study design should be used to support a submission for a particular device.*** This guidance document describes different study design principles relevant to the development of medical device clinical studies that can be used to fulfill pre-market clinical data requirements. ***This guidance is not intended to provide a comprehensive tutorial on the best clinical and statistical practices for investigational medical device studies.***

(Page 4) (Emphasis added).

Moreover, notwithstanding PCAST's assertion to the contrary, the FDA Report on the Adaptive Design for Medical Device Clinical Studies provides that "under certain circumstances, a number of scientifically valid changes to the study design ***can be*** entertained ***even if*** they are ***not*** preplanned" (p. 26). (Emphasis added).

As an aside, PCAST claims that "[i]n the design of clinical trials, FDA requires that criteria for analysis must be pre- specified and notes that ***post hoc changes to the analysis compromise the validity of the study.***" (Emphasis added). This assertion is apparently based on a statement contained in the FDA's guidelines concerning "Adaptive Designs for Medical

Device Clinical Studies." However, PCAST takes the actual quote in the guidelines out of context and rephrases it to suit its own purposes. The actual language states:

> Any ***change or revision to a study design*** is post hoc and not adaptive if it is based on unplanned findings from an interim (or final) analysis in a study where the blind (mask) of outcomes by treatment groups has been broken (even if only the coded treatment group outcomes). Such modifications generally would endanger the scientific validity of the study since the false positive rate is not controlled and there is a strong possibility of operational bias.

(FDA Adaptive Design Guidelines, Page 9).

PCAST's alteration of the text is significant because it substitutes *its* chosen word, "analysis," for the *actual* words "study design" and then utilizes that alteration to attack the legitimacy of the Miami-Dade latent fingerprint study findings, which showed a very low false positive rate (PCAST Report, p. 95). Regarding those findings, PCAST states: (Note: The paper observes that in 35 of the erroneous identifications the participants appeared to have made a clerical error, but the authors could not determine this with certainty.) ***"In validation studies, it is inappropriate to exclude errors in a post hoc manner*** (see Box 4)." (Emphasis added).

If anything, Miami-Dade's *post hoc* conjectures on the possible reasons for the number of incorrect results was an *analysis* of the results, *not* a *post hoc* "change or revision to a *study design*," as discouraged by the FDA document. However, PCAST uses Miami-Dade's *post hoc* determination that 35/42 false positive results were probably the result of clerical errors, in conjunction with the misquoted text from the FDA's Adaptive Designs document, as reasons to reject a much lower false positive rate in favor of a much higher one (false positive upper bound of 1 error in 18 cases). Thus, in addition to incorrectly claiming that the FDA *requires* that certain criteria must be satisfied in the design and performance of clinical trials, PCAST also incorrectly paraphrased one of its chosen sources to suit its own purposes.

The third problem with these criteria is that they are based upon a faulty analogy between forensic feature comparison methods and the clinical study of "medical devices" conducted on laboratory animals, human subjects, and non-clinical *in-vitro* studies (Design Considerations guidelines, p. 9). PCAST fails to even attempt to justify this analogy, apparently hoping that the reader will simply accept the comparison as valid at face value and question no further. The fact is that there are substantial differences between clinical trials designed to measure the impact of a "medical device" on the health, safety, and well-being of an animal, a human, or an unborn child, and the ability of forensic examiners to correctly recognize, analyze, compare, and distinguish between questioned and known features. PCAST, however, simply assumes that these different subjects, aims, and objectives should have no impact on the appropriate design and experimental criteria to be utilized to conduct these very different types of studies. Because PCAST has not even attempted to justify its strained analogy between clinical trial designs and forensic feature comparison validation studies, it

has failed to prove the applicability of its validation criteria. As such, the analogy must be rejected as faulty.

The fourth problem with these criteria is that their rigid formulation and mandatory nature is not only inconsistent with the non-mandatory nature of the FDA's guidance concerning clinical trial designs and experiments, but is also inconsistent with the definition of "Valid Scientific Evidence" contained in the FDA's own Design Considerations document. That definition states:

> Valid scientific evidence is evidence from well- controlled investigations, partially controlled studies, studies and objective trials without matched controls, well-documented case histories conducted by qualified experts, and reports of significant human experience with a marketed device, from which it can fairly and responsibly be concluded by qualified experts that there is reasonable assurance of the safety and effectiveness of a device under its conditions of use. The evidence required may vary according to the characteristics of the device, its conditions of use, the existence and adequacy of warnings and other restrictions, and the extent of experience with its use. Isolated case reports, random experience, reports lacking sufficient details to permit scientific evaluation, and unsubstantiated opinions are not regarded as valid scientific evidence to show safety or effectiveness. Such information may be considered, however, in identifying a device the safety and effectiveness of which is questionable.

(Design Considerations guidelines, p. 9 (quoting 21 CFR 860.7(c)(1)).

It is clear that the rigid and mandatory set of criteria that PCAST would impose upon forensic validation studies finds no support in this definition — the very benchmark against which scientifically valid clinical trials are measured. Rather, as the definition notes, the determination that scientific evidence is valid may be derived from a variety of study types and designs. These include "well-controlled investigations;" "partially controlled studies;" "studies and objective trials without matched controls;" "well-documented case histories conducted by qualified experts;" and "reports of significant human experience."

Each of these types of studies has been performed by those who have conducted research on the various feature comparison methods criticized by PCAST. In addition, a substantial amount of empirical evidence has been generated by these studies. PCAST, however, provides no explanation why the data generated by the types of studies listed in the federal definition of "valid scientific evidence" — which are apparently *sufficient* to support valid clinical trial designs and evaluations — are nevertheless *insufficient* to support the validation of forensic feature comparison methods. PCAST's failure to explain this discrepant treatment further calls into question both its stated criteria and the analogy that it offers between clinical trials and feature comparison method validation.

The fifth problem with these criteria is that some of them are unsupported by a single

citation to extant scientific literature on the topic of scientific validation. Absent a single citation that would individually or collectively support their inclusion in PCAST's list of criteria, one is left to assume that some of these elements are simply the original creation of the Working Group. This suspicion is augmented by the fact that PCAST attempts to bolster the gravitas of its Report, and in turn the questioned criteria, by a logically fallacious appeal to its own authority.

On page 144 of its Report, PCAST asserts that "from a *scientific* standpoint, ***subsequent events*** have indeed undermined the continuing validity of conclusions that were not based on appropriate empirical evidence." (Emphasis added). Incredibly, PCAST attempts to fortify this conclusion by bootstrapping *its own Report* as one of these "subsequent events," claiming, "the scientific review in this report by PCAST, ***the leading scientific advisory body established by the Executive Branch***, finding that some forensic feature-comparison methods lack foundational validity" is one of those events (PCAST Report, p. 144). This determination was, of course, based upon PCAST's concurrent conclusion that the feature comparison methods examined in its Report did not fully meet each of its newly-minted validation criteria. Thus, rather than citing established scientific literature in support of its criteria, PCAST instead appeals to its readers' acceptance of a logical fallacy as a substitute for scientific authority. This maneuver is not persuasive. It wholly fails to support the scientific *necessity* that *each* of its listed factors be satisfied as conditions precedent to establishing the foundational validity of the feature comparison methods examined.

The sixth problem with these criteria is PCAST's apparent failure to consider (or lack of concern about) the importance of establishing the external validity of validation studies that may incorporate its mandatory criteria. In science, "external validity" concerns whether or not the findings derived from scientific research studies can be extrapolated to individual events and instances. In other words, it is the extent to which the results of a study can be generalized to other situations and to other people.

This is important because, in its Report, PCAST urges that one "key criteria for validity as applied" is that "[t]he forensic examiner should [for both reports and testimony] ***report the overall false positive rate and sensitivity*** for the method ***established in the studies of foundational validity*** and should demonstrate that ***the samples used in the foundational studies are relevant to the facts of the case***." (PCAST Report, p. 56) (Emphasis added). Furthermore, PCAST asserts, "Since empirical measurements are based on a limited number of samples, SEN [sensitivity] and FPR [false positive rate] cannot be measured exactly, but only estimated. Because of the finite sample sizes, the maximum likelihood estimates thus do not tell the whole story. Rather, it is necessary and appropriate to quote confidence bounds within which SEN, and FPR, are highly likely to lie" (PCAST Report, 152). To that end, the Report advocates that:

> ***Because one should be primarily concerned about overestimating SEN or underestimating FPR, it is appropriate to use a one-sided confidence bound***. By convention, a confidence level of 95 percent is most widely

> used—meaning that there is a 5 percent chance the true value exceeds the bound. Upper 95 percent one-sided confidence bounds should thus be used for assessing the error rates and the associated quantities that characterize forensic feature matching methods. ***(The use of lower values may rightly be viewed with suspicion as an attempt at obfuscation.)***[1]

Applying these principles to latent print casework, for example, it is PCAST's position that an examiner should report the following to a jury:

> [T]hat (1) only two properly designed studies of the accuracy of latent fingerprint analysis have been conducted and (2) these studies found false positive rates that could be as high as 1 in 306 [only the upper bound value] in one study and 1 in 18 [only the upper bound value] in the other study. This would appropriately inform jurors that errors occur at detectable frequencies, allowing them to weigh the probative value of the evidence.

(PCAST Report, p. 96).

It is apparently PCAST's position that not only should its validation criteria be used by foundational feature comparison studies without any attempt to establish the external validity of findings derived from those studies when applied to casework and testimony, but *also* that the jury should *only* be informed of the upper bound confidence values derived from studies whose external validity has not been established. The former position promotes the generalization of data to specific cases when the scientific legitimacy of those inferences has not been established. The latter position is little more than thinly-veiled partisanship masquerading as science.

Accordingly, because PCAST has failed to demonstrate that its requisite criteria for establishing foundational validity will provide externally valid rates for false positives and method sensitivity when applied to individual cases, its insistence that validation studies "must satisfy" these criteria is unpersuasive.

Conclusion

In Chapter 9, *Actions to Ensure Scientific Validity in Forensic Science: Recommendations to the Judiciary,* the PCAST Report calls for the courts to re-examine court decisions that admitted forensic feature-comparison methods and overrule those decisions. PCAST claims the courts

[1] This statement has already received academic criticism for its biased formulation. In his Forensic Science, Statistics & the Law blog, Professor David Kaye states, "I have to say that this paragraph seems to contradict the ideal of a forensic scientist who does not take sides." Professor Kaye continues by observing, "It is fair to say that the *exclusive* use of lower values — instead of both upper and lower values — may rightly be viewed with suspicion as an attempt at obfuscation. It is equally fair to say that the *exclusive* use of upper values also may rightly be viewed with suspicion as an attempt at obfuscation." *PCAST's Sampling Errors*, David Kaye, Forensic Science, Statistics & the Law, October 24, 2016 (original emphasis). Available at: http://for-sci-law.blogspot.com/2016/10/pcasts-sampling-errors.html.

should give deference to the 2009 NRC report from the National Academy of Sciences and "the scientific review in this report by PCAST."

PCAST should take note that the conclusions of the NRC report have already been addressed in *United States v. Rose*, 672 F. Supp. 2d 723 (D. Md. 2009) where the Federal Judge quoted from Judge Harry Edwards, who co-chaired the project, "made it clear that nothing in the Report was intended to answer the question whether forensic evidence in a particular case is admissible under applicable law."

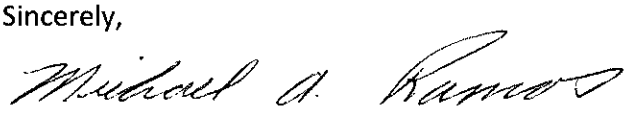
Meanwhile in the short period of time since the release of the PCAST report, it has been generally rejected by the Courts as having no relevance to the issue of admissibility feature comparison evidence (See Appendix A).

In conclusion, instead of providing PCAST with a list of "***appropriately designed, research studies***" establishing the foundational validity of contemporarily practiced feature comparison methods (per its own definition of "appropriate design"), the burden should be on PCAST to *first* provide the forensic community with information in support of the following requests:

> (1) Provide your *own* list of relevant scientific authorities that specifically justify the criteria that you claim forensic feature comparison methods "must satisfy." (PCAST Report, pp. 52-52); (2)Specifically explain how and why these authorities have direct applicability to establishing the foundational validity of the questioned feature comparison methods; and (3) Explain why the satisfaction of each and every one of the listed criteria is a *sine qua non* for establishing the foundational validity of the feature comparison methods examined in your Report, despite the fact that these same criteria are *not* mandatory in other realms of scientific validation — such as your chosen analogy — clinical trial design and experimentation.

Only *after* PCAST provides the forensic science community with sufficient information and convincing reasons which establish the scientific standing of its validation criteria should the forensic science community bear the burden of providing PCAST with studies that can be assessed against the elements of its currently disputed concept of an "appropriately designed" research study. The responsible approach would be for the PCAST to withdraw the September 2016 publication on the grounds extensive research was not conducted prior to the document being published and the PCAST is now soliciting citations for additional research.

Sincerely,



Michael A. Ramos
President
National District Attorneys Association

Appendix A
Recent Court Decisions Since Publication of the PCAST Report

Commonwealth v. Legore, the Superior Court for Massachusetts refused to reverse the precedents admitting firearms ballistic comparison. The court said, "After a non-evidentiary hearing and argument, and upon review of the PCAST report, there is no basis to disturb settled law permitting a properly qualified firearms expert from offering opinion evidence under Rule 702 relating to a comparison and match...."

U.S. v. Chester, United States District Court for the Northern District of Illinois Eastern Division, "In short, the PCAST report does not undermine the general reliability of firearm toolmark analysis or require exclusion of the proffered opinions in this case."

People v. Michael Robinson, defense attorneys argued that TrueAllele was "novel," based on a 2016 report from the President's Council of Advisors on Science and Technology (PCAST). This report advised DNA technology limitations without citing scientific support. Since the defense did not present evidence supporting PCAST findings, the prosecution reasoned there should not be a hearing. The judge found according to Pennsylvania precedent, not PCAST report. In 2012, a Superior Court ruled TrueAllele science admissible.

Minnesota v. Yellow, defense attorneys argued the PCAST was an important development, and should encourage the court to reconsider its prior decision on the foundational reliability of complex DNA mixtures. The Court found that the opinions met the standard for foundational reliability, and nothing in the PCAST Report changes that finding.

Dear PCAST Committee, This email is in response to your request for additional information on foundational validity in forensic sciences. Firearms identification has been practiced for over 100 years in the United States and two research papers are cited for your review (there are obviously many, many others as well). You asked for 'firearms analysis to associate ammunition with an individual gun'. The papers are titled:

Evaluation of GLOCK 9 mm Firing Pin Aperture Shear Mark Individuality Based On 1,632 Different Pistols by Traditional Pattern Matching and IBIS Pattern Recognition

James E. Hamby,[1] Ph.D.; Stephen Norris,[2] B.S.; and Nicholas D.K. Petraco,[3,4] Ph.D.

(Published in the January 2016 Issue of the Journal of Forensic Sciences)

*(This is an ongoing research project and is approaching over **3,000** different* pistols, *manufactured during a 25 year period in both Austria and the United States. Test fired ammunition components represents some 24 different headstamp, primer and cartridge case compositions.)*

The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries

James Hamby, Ph.D., David Brundage, M.S., and James Thorpe, Ph.D.

Published in the AFTE Journal, Volume 41, Number 2, Fall 2009

(This is an ongoing research project and to date has some 686 participants – eleven of whom used sort type of ballistics imaging (many of whom are not firearms examiners) – from 32 countries. Although it was referred to as a 'suduko' type of event, it is obviously a complex examination with an incredibly low error rate. The researchers recognize that it is a closed set (sampling without replacement) but also represents the ability to individualize bullets fired from consecutively manufactured barrels)

Thank you for your consideration to this email. James E. Hamby, Ph.D.

From: [redacted]
To: [redacted]
Subject: Toolmarks
Date: Saturday, December 3, 2016 11:46:58 AM

---

https://afte.org/uploads/documents/postion-pcast-2015.pdf

To whom it may concern,

AFTE has published many articles related to the study of Toolmarks when it comes to firearms and the evaluation of evidence for examination. The above link in my opinion provides articles with evidence based research into the science with proven results. I, myself , have participated in a study that included a "ten barrel test" . The reproduceable markings can be quantified by a properly trained examiner whose experience and knowledge results from similarly based programs that qualify those as a Toolmarks examiner by AFTE standards. A majority of us today, especially independents such as myself, continue to not only keep up to date with current journal articles from resources such as AFTE, but continue to participate in blind examinations such as those from private companies as CTS to remain proficient. At the conclusion of their studies, we are provided with results from the organization which reflect an accurate error rate which is generally consistent with the field.
Not every investigation into ballistics evaluation of recovered evidence is clear cut, but with the proper training and experience I feel in my opinion the scientific conclusions that I personally have reached were those based on the scientific principles and proven foundation of Toolmarks comparisons. These studies are repeated often and documented as such in the various scholarly journals. The science as a whole welcomes inquiries and is constantly critically analyzing any new developments in the field from a host of sources

Merry Christmas and Happy New Year

Bruno R.Valenti

Sent from my iPhone

Organization of Scientific Area Committees (OSAC)
Firearms and Toolmarks Subcommittee

Response to the President's Council of Advisors on Science and Technology (PCAST) Call for Additional References Regarding its Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods"

14 December 2016

The Organization of Scientific Area Committees (OSAC)[1] Firearms and Toolmarks Subcommittee is composed of sixteen forensic practitioners with a combined 307 years of forensic science experience. The practitioners are drawn from federal, state, county, local and private laboratories from across the country. Additionally, the subcommittee includes four non-practitioners with backgrounds in metrology, statistics, and computer science. The subcommittee's composition meets OSAC's goals of diversity of both forensic practitioners and non-practitioners. Given the responsibility of the subcommittee for informing the process of developing standards and guidelines for the forensic discipline of firearm and toolmark identification, we feel it necessary to respond to the report published by the President's Council of Advisors on Science and Technology (PCAST) and the subsequent Request for Information (RFI) distributed by PCAST co-chair Dr. Eric Lander on December 2, 2016.

The PCAST report addresses numerous subjects and seven disciplines of forensic science. We will limit our response to those portions addressing firearm and toolmark identification. We disagree with PCAST's conclusion that "...firearms analysis currently falls short of the criteria for foundational validity, because there is only a single appropriately designed study to measure validity and estimate reliability." This response will outline why we find PCAST's analysis to be inaccurate.

[1] The purpose of the Organization of Scientific Area Committees (OSAC) is "...to strengthen the nation's use of forensic science by providing technical leadership necessary to facilitate the development and promulgation of consensus-based documentary standards and guidelines for forensic science, promoting standards and guidelines that are fit-for-purpose and based on sound scientific principles, promoting the use of OSAC standards and guidelines by accreditation and certification bodies, and establishing and maintaining working relationships with other similar organizations." https://www.nist.gov/topics/forensic-science/about-osac

1 Black-Box (Validation) Study Analysis

PCAST analyzed nine firearm black-box studies and concluded that firearms identification "falls short of the criteria for foundational validity."[2] We disagree with their position because it ignores critical details within each study and their review falls short in understanding the research value these studies provide when considered in totality. Additionally, other validation studies have been performed that were not addressed by PCAST.[3,4,5,6,7,8]

1.1 Introduction

Black-box studies (a common type of validation study) use ground truth to evaluate the soundness and accuracy of examinations. PCAST required that a validation study be of "black-box" design and that samples be examined completely independently of each other. PCAST set the following criteria for determining if a forensic science discipline is scientifically valid: 1) at least two black-box studies that allow for the calculation of a False Positive Error Rate (FPR) and 2) an error rate less than 5%[9]. There is no reference or justification to support that this is a generally-accepted standard.

The studies examined by PCAST were categorized into four different types: "within-set," "set-to-set," "partly open set," and "independent/open." Within these categories, PCAST examined nine validation studies and discounted the data from eight due to test design. PCAST also made errors when summarizing these studies. They did not accurately count the number of responses, or left data out, from four of the nine validation studies used for their analysis. A summary of the errors can be found in Appendix A.

---

[2] PCAST Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods"," (September, 2016) Finding 6, pg 112.

[3] Lyons, D. J. "The Identification of Consecutively Manufactured Extractors." *AFTE Journal*, Vol. 41, No. 3 (2009): 246-256.

[4] Bunch, S. G., and D. Murphy. "A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases." *AFTE Journal*, Vol. 35, No. 2 (2003): 201-203.

[5] Mayland, B. and C. Tucker. "Validation of Obturation Marks in Consecutively Reamed Chambers." *AFTE Journal*, Vol. 44, No. 2 (2012): 167-169.

[6] Fadul, T. G. "An Empirical Study to Evaluate the Repeatability and Uniqueness of Striations/Impressions Imparted on Consecutively Manufactured Glock EBIS Gun Barrels." *AFTE Journal*, Vol. 43, No 1 (2011): 37-44.

[7] Cazes, M. and J. Goudeau. "Validation Study Results from Hi-Point Consecutively Manufactured Slides." *AFTE Journal*, Vol. 45, No. 2 (2013): 175-177.

[8] A listing and summary of additional supportive research, and validation studies pertaining to non-firearm toolmarks, can be found in the SWGGUN Admissibility Resource Kit (ARK). https://afte.org/resources/swggun-ark/testability-of-the-scientific-principle

[9] PCAST Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods"," (September, 2016) "Finding 6", pp 112, Appendix A, pg 152.

Below we summarize PCAST's analysis and why we disagree with their findings.

1.2 Within-set Studies

PCAST summarized two "within-set" validation studies.[10,11] The PCAST committee could not calculate a False Positive Error Rate (FPR) using these studies, so they did not use them to measure the validity of firearm and toolmark identification.

The dismissal of these studies does not accurately reflect the scientific value of the research. A total of 1037 different-source comparisons were performed. No false identifications or false eliminations were reported by any of the participants. Therefore, these two studies provide empirical and independent support that the overall error rate for firearm and toolmark identification is low, despite the inability to calculate a false positive error rate.

1.3 Set-to-Set Comparison/Closed Set Studies

PCAST summarized four "closed set" studies.[12,13,14,15] PCAST is critical of these test designs because each comparison is not independent of the others. The assumption is that examiners may be able to deconstruct the test design, and PCAST likens this to the same logic as solving a "Sudoku" puzzle.[16] The analogy used by PCAST misrepresents the challenge posed by these tests. First, three of the studies (Brundage et al., Hamby et al., Fadul et al.) used consecutively manufactured firearms. Consecutively manufactured firearms have been shown to have the potential for subclass characteristics, which are toolmarks that sometimes carry over, with very

---

[10] Smith, E. "Cartridge case and bullet comparison validation study with firearms submitted in casework." *AFTE Journal*, Vol. 37, No. 2 (2005): 130-5. There were a total of 16 same-source comparisons and 704 different-source comparisons in this study. 13 of the 16 same-source comparisons were correctly identified and 3 were inconclusive. There were no false identifications or false eliminations reported.

[11] DeFrance, C.S., and M.D. Van Arsdale. "Validation study of electrochemical rifling." *AFTE Journal*, Vol. 35, No. 1 (2003): 35-7. There were a total of 45 same-source comparisons and 333 different-source comparisons. 42 of the 45 same-source comparisons were correctly identified and 3 were inconclusive. There were no false identifications or false eliminations.

[12] Stroman, A. "Empirically determined frequency of error in cartridge case examinations using a declared double-blind format." *AFTE Journal,* Vol. 46, No. 2 (2014):157-175.

[13] Brundage, D.J. "The identification of consecutively rifled gun barrels." *AFTE Journal*, Vol. 30, No. 3 (1998): 438-44.

[14] Fadul, T.G., Hernandez, G.A., Stoiloff, S., and S. Gulati. "An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides." *AFTE Journal.* Vol. 45, No. 4 (2013): 376-93.

[15] Hamby, J.E., Brundage, D.J., and J.W. Thorpe. "The identification of bullets fired from 10 consecutively rifled 9mm Ruger pistol barrels: a research project involving 507 participants from 20 countries." *AFTE Journal,* Vol. 41, No. 2 (2009): 99-110.

[16] PCAST Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" (September, 2016) Section 5.5, pp 106. PCAST was quoting Jeff Salyards, Director of the Defense Forensic Science Center.

little change or variation, from one machined part to the next on the same production line.[17,18,19] Qualified examiners are able to recognize these marks so as not to use them for conclusions of identification. Though consecutively manufactured firearms are not likely to be encountered in actual casework, the authors used them in an attempt to create a worst-case scenario (i.e. potential best known non-matches). Additionally, each test used *more* questioned samples than knowns (15 questioned samples from 10 consecutively manufactured firearms). Therefore, taking these tests was not as simple as figuring out a few of the correct answers and then deducing the rest. Since these tests used consecutively manufactured samples, it was just as important to know if examiners could correctly identify samples as it was to know if samples were falsely identified. This is the reason at least one true match was provided with each questioned cartridge case.

Another study discounted by PCAST was conducted by Stroman et al. This validation study used cartridge cases that had been fired in Smith & Wesson pistols. While this study did not use consecutively-manufactured samples, the firearms were the same make and model and had documented subclass characteristics on the firearms' ejectors. Again, these are potentially difficult samples and provide the opportunity for false positive errors, yet none were observed.

In each of these four studies, the authors attempted to create tests with potentially challenging samples. Each of these studies provide insight into the overall error rate (see Appendix A for more details about each study). The fact that few false positive errors occur is strong evidence in support of the discipline of firearm and toolmark identification. These studies present evidence that firearm and toolmark examiners can reliably and accurately associate questioned toolmarks to the correct source tool. Though the test design does not fit the model proposed by PCAST, these studies present valuable performance estimates and should not be disregarded. When viewed collectively, these studies are independent of each other and show a low overall error rate among the tested examiners. This provides strong support for the overall validity of firearm and toolmark identification.

---

[17] Weller, T.J., Zheng, X.A., Thompson, R.M., and F. Tulleners. "Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides." *Journal of Forensic Sciences,* Vol. 57, No. 4 (2012): 912-17. This study has documented subclass characteristics among the 10 consecutively manufactured pistol slides. An eleventh pistol slide, that was not part of the consecutive batch, no longer has the same subclass toolmarks.

[18] Miller J., Beach G. "Toolmarks: Examining The Possibility of Subclass Characteristics" *AFTE Journal*, Vol 32, No 4: 296-345.

[19] Subclass characteristics are features that may be produced during manufacture that are consistent among items fabricated by the same tool in the same approximate state of wear. These features are not determined prior to manufacture and are more restrictive than class characteristics. AFTE Glossary, 6th Edition.

1.4 Partly Open Set

PCAST summarized another validation study and categorized it as "partly open."[20] We would like to highlight the fact that this study also uses consecutively manufactured samples and, as described above, provides examiners with test samples which are most likely to have similar toolmarks since the firearms used to create them were sequentially manufactured with the same tools.

PCAST's statistical analysis of this report focused solely on two unknowns that had no matching known. This analysis is incomplete, and differs from the analysis used by PCAST in the "set-to-set/closed set" and "open set" studies where all "conclusive" responses were used to calculate the False Positive Error Rate.

The authors' reported error rate (0.7%) was low and this study provides an additional independent study establishing that firearm and toolmark examiners can accurately associate questioned toolmarks to the correct source tool.

1.5 Open Set

PCAST summarized another validation study and categorized it as "open."[21]

Each test taker in this study was instructed to work independently and not collaborate with other test takers. These instructions negate an important quality assurance step used in most accredited forensic laboratories: the peer review process known as verification[22]. Verification is a reevaluation of a comparison by another qualified examiner to ensure there is sufficient data to support the conclusion. Many laboratories accomplish this by direct reexamination of the evidence, while others use representative photographs of sufficient quality for the verification step. The errors reported in this paper may have been caught if verification were allowed. This suggests the true false positive error rate may be lower than calculated in this study. We would like to highlight that Baldwin et al. discusses this point in their study (emphasis added):

> "This finding does not mean that 1% of the time each examiner will make a false-positive error. Nor does it mean that 1% of the time laboratories or agencies would report false positives, **since**

[20] Fadul, T.G., Hernandez, G.A., Stoiloff, S., and S. Gulati. "An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing consecutively manufactured Glock EBIS barrels with the same EBIS pattern." National Institute of Justice Grant #2010-DN-BX-K269, December 2013.

[21] Baldwin, D.P., Bajic, S.J., Morris, M., and D. Zamzow. "A study of false-positive and false-negative error rates in cartridge case comparisons." Ames Laboratory, USDOE, Technical Report #IS-5207 (2014) afte.org/uploads/documents/swggun-false- postive-false-negative-usdoe.pdf.

[22] In the other validation studies discussed above, verification was also unlikely because test takers were not to collaborate with other test takers.

> **this study did not include standard or existing quality assurance procedures, such as peer review or blind reanalysis.** What this result does suggest is that quality assurance is extremely important in firearms analysis and that an effective QA system must include the means to identify and correct issues with sufficient monitoring, proficiency testing, and checking in order to find false-positive errors that may be occurring at or below the rates observed in this study. "[23]

It should be noted that PCAST used the data from the study to recalculate a false positive error rate by using only exclusion conclusions and omitting the inconclusive results. This resulted in a rise in the calculated error rate from 1.01% to 1.5%. The different error rates provide different answers for different questions: The Baldwin et al. error rate estimates how often non-matching cartridge cases are falsely identified, while PCAST's error rate estimates the proportion of definitive (i.e. not inconclusive) results that are incorrect when non-matching cartridge cases are examined.

Baldwin et al. provide a discussion about inconclusive results (emphasis added):[24]

> "If the examiner does not find sufficient matching detail to uniquely identify a common source for the known and questioned samples, and there are no class characteristics such as caliber that would preclude the cases as having been fired from the same-source firearm, **a finding of inconclusive is an appropriate answer (and not counted as an error or as a non-answer in this study).** The underlying rationale for this finding of inconclusive is that the examiner is unable to locate sufficient corresponding individual characteristics to either include or exclude an exhibit as having been fired in a particular firearm and the possible reasons are numerous as to why insufficient marks exist. As is determined in this study, there are also a significant number of times that the firearm fails to make clear and reproducible marks (which very well might have happened for a questioned case)."

Baldwin et al. found the rate of poor quality mark production to be 2.3% (+/- 1.4%). This rate is double the calculated false positive error rate. This provides support for the use of inconclusive results in the calculation of error rates.

We would like to highlight the fact that the Baldwin study found "all but two of the 22 false identification calls were made by five of 218 examiners."[25] This indicates when errors do occur, they may be committed by the same few examiners. This supports the need for rigorous

[23] Baldwin et al. Pg 18.
[24] Baldwin et al. Pg 6
[25] Baldwin et al. Pg 16.

training, periodic proficiency testing, continuing education and thorough laboratory quality control measures.

1.6 Smith et al. Study

The final validation study examined by PCAST was the Smith et al. study, in which the authors created a test that mimics casework[26]. PCAST concluded this study was insufficient to test the validity of firearm identification:

> "While interesting, the paper clearly is not a black-box study to assess the reliability of firearms analysis to associate ammunition with a particular gun, and its results cannot be compared to previous studies."[27]

PCAST recognizes the study as being new and novel. We disagree with their observation that since the study is not a "black-box" design then the study does not provide support for the validity of firearm identification. In the test design that PCAST requires, test takers examine only one questioned sample at a time, independent of other questioned samples. While we understand this test design allows for easier statistical analysis, one to one comparisons are not an accurate representation of actual casework. A typical examination for a firearm examiner entails opening a package of evidence with dozens of items and attempting to associate or disassociate the items. This study tested that process by forcing examiners to make all of the typical decisions they would make in casework, rather than conducting a series of examinations on isolated pairs of specimens. The test takers were presented with bullets and cartridge cases of various ammunition types, and asked to perform both class and individual characteristic evaluations. They were not given any information about the source of any of the items.

Test takers were faced with a real-world scenario and performed very well. Although not stated in the PCAST footnote referencing this article, the overall error rate for this study was 0.303%.

1.7 Conclusions

PCAST reviewed nine validation studies and through their criteria, elected to discount eight of those studies. Two of those disregarded studies (the "within-set" design) had no false positive results. Five of the disregarded studies had very few false positives (see Appendix A) and the last study (which attempted to replicate casework) found a low overall error rate (0.303%).

---

[26] Smith, T., Smith, G.A., Snipes, J.B. "A Validation Study of The Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework." Journal of Forensic Sciences, Vol. 61, No. 4: 939-946

[27] PCAST Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" (September, 2016), footnote #335.

When PCAST set criteria for the validity of a forensic science discipline, they chose an arbitrary threshold of having at least two black box studies. The black-box test design favored by PCAST requires that each questioned sample be examined independently from each other. Examiners are not faced with completely independent examinations when they analyze evidence in a case. It is not realistic, if trying to replicate casework, to have fifteen or twenty individual sets of comparisons, each of which is made independent of each other. The PCAST-proposed design may make sense from a purely statistical standpoint, but does not simulate the practical task of an examiner performing casework. The OSAC subcommittee believes that various types of tests are valuable and can provide meaningful information regarding the potential error rates[28].

2.0 Subjective and Objective Methods

PCAST defines objective feature comparison methods as "methods consisting of procedures that are each defined with enough standardized and quantifiable detail that they can be performed by either an automated system or human examiners exercising *little or no judgment*" (emphasis added). PCAST defines subjective methods as "methods including key procedures that involve *significant human judgment*[29]" (emphasis added).

In fact, all disciplines, including firearm and toolmark identification, require some human judgment or interpretation of results. Implementation of more objective techniques may make those interpretations easier, but judgment will still be required.

We agree, however, with the goal of continuing to research and implement more objective analytical methods. One of our subcommittee's task groups is writing standards that will assist industry and crime laboratories with the validation and implementation of new technology. Additionally, there is a growing body of research using three-dimensional instrumentation and advanced machine-learning algorithms to compare toolmarks. The research fails to disprove the foundational premise of firearm and toolmark identification: that fired ammunition components can be associated to (or eliminated from) the originating firearm through the comparison of microscopic toolmarks. In fact, the recent research provides strong objective

---

[28] Different test designs estimate different error rates. For example: when examining evidence from an officer involved shooting where each officer admits to firing their firearm: error rates based on data from "set to set/closed-set" studies may be more appropriate while the Smith et. al. study may provide a better estimate for an examination of numerous items with no questioned firearm. All of these studies have the potential to provide a relevant error rate estimates and the "true" error rate may not be the same for each situation.

[29] PCAST Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" (September 2016) Section 4.1, pp. 46-47.

support for this premise. The PCAST committee was provided with 25 citations by our subcommittee documenting this work; however, their report only cites two studies.

3.0 AFTE Theory of Identification is Circular Logic

PCAST states that the AFTE Theory of Identification is circular logic. PCAST's summary of the theory makes it sound circular:

> "It declares that an examiner may state that two toolmarks have a "common origin" when their features are in "sufficient agreement." It then defines "sufficient agreement" as occurring when the examiner considers it a "practical impossibility" that the toolmarks have different origins."[30]

The PCAST Report makes the AFTE Theory sound circular by ignoring the basis for "sufficient agreement." This is based on a misunderstanding of what constitutes "sufficient agreement." They claim it is an arbitrary point at which the examiner considers it a "practical impossibility." PCAST seems to believe that this "practical impossibility" is arbitrarily decided by the examiner, thus making the theory sound circular. This is incorrect. The sufficient agreement threshold is exhibited when the amount of agreement is greater than best known non-matches established by the community and conveyed to each examiner through a lengthy and extensive training program. That is, it is not an arbitrary point. In fact, by definition, no non-matches can ever have more similarity than the sufficient agreement point. When the basis for the ground truth is included, the AFTE Theory is not circular.

4.0 Focus on Training and Experience Rather Than Empirical Demonstration of Accuracy

PCAST quote:

> "Many practitioners hold an honest belief that they are able to make accurate judgments about identification based on their training and experience.[31]"

In all professions, proper training and experience is critical. Firearm and toolmark identification is like other applied sciences (e.g. medicine, engineering) that require training to become proficient and experience to further refine and maintain that proficiency. There is only so much that textbooks can teach, and structured training (like residency for physicians) is a critical aspect of developing proficiency. It is through rigorous training that examiners develop their criteria for what constitutes an elimination, an identification, or an inconclusive result. They

---

[30] PCAST Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" (September 2016) Section 4.7, pp. 60.

[31] PCAST Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" (September 2016) Section 4.7, pp 60-61.

learn and understand the differences in microscopic agreement between toolmarks created by the same source (a known match) and toolmarks created by different sources (a known non-match) and how that understanding factors into any conclusion of elimination, inconclusive, or identification. Examiners do not memorize all patterns that have been observed, as suggested in the PCAST report.

5.0 Conclusion

The Firearms and Toolmarks Subcommittee of OSAC fundamentally disagrees with the conclusions regarding the firearm and toolmark identification discipline presented in the PCAST report. Four major points have been put forth in this response. First, we disagree with the premise that a structured black-box study is the only useful way to gain insight into both the foundations of firearm and toolmark identification and examiner error rates. Taken collectively, the published studies support the underlying principles of firearm and toolmark examination and the fact that examiner error rates are quite low. PCAST's critique of these studies included several misunderstandings. Second, PCAST's dismissal of methods employing a subjective component discounts the core scientific methods that have been used for hundreds of years. Third, PCAST misunderstands and misquotes the AFTE Theory of Identification. PCAST's summary of the AFTE Theory of Identification leaves out important provisions. Fourth, PCAST minimizes the value of training and experience. The training received by firearm examiners includes both subjective and objective components and is comparable to the domain-specific rigor of other applied scientific fields.

We do not agree that firearm identification "...falls short of the criteria for foundational validity." However, we do agree that a hallmark of any scientific endeavor is ongoing research and technology development. Indeed, our subcommittee, which is tasked with writing standards and providing guidance to the profession, would not exist if it was believed that the field of firearm identification is flawless and requires no improvement. As such, we are hopeful that the path forward from the PCAST report is a renewed commitment to research in the forensic sciences, continued testing of foundational principles, and a more robust collaboration between the academic and forensic practitioner communities.

Appendix A
Errors and Omissions in PCAST Summaries of Firearms and Toolmarks Validation Studies

PCAST incorrectly summarized four of the nine validation studies used in their analysis of firearm and toolmark identification. For clarity, we first repeat some of the terms used by PCAST to illustrate how they (and we) calculated these error rates.

> "The results of a given empirical study can be summarized by four values: the number of occurrences in the study of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN)"[32]

PCAST used the following formula to calculate the "maximum likelihood estimate of FPR": FP/(FP+TN).[33] For those unfamiliar with statistics, we recalculate the FPR for the Baldwin et al. study. There were a total of 2178 different-source comparisons performed: 1421 were declared elimination, 735 were reported as inconclusive, and there were 22 false positives reported. PCAST did not use inconclusive results in their statistical treatment (as we discussed in Section 1.5). Therefore, PCAST's FPR calculation for the Baldwin et al. study is: FPR= 22/(1421+22). This equals 0.015, or 1.5%. Conversely, recognizing that inconclusive results are appropriate[34], Baldwin, et al. included inconclusive results in their calculations, as follows: FPR= 22/(1421+735+22). This equals 0.010, or 1.0%.[35]

For the "set-to-set/closed" studies, PCAST used correct identifications in lieu of using true negatives[36]. PCAST does not explain or justify why they did this. The error rates reported by PCAST for the "set-to-set/closed" studies found in Table 2 on page 111 of the PCAST report are not false positive error rates and should not be reported as such.

Below we summarize the errors made by PCAST in their assessment of four of the nine studies.

---

[32] PCAST Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods", Appendix A, pg 152.

[33] PCAST Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods", Appendix A, pg 152.

[34] Baldwin et al., pg 6.

[35] Baldwin et al., pg 16.

[36] See footnote 327 of PCAST report: "Of the 10,230 answers returned across the three studies, there were there were 10,205 correct assignments, 23 inconclusive examinations and 2 false positives."

Brundage Study

The PCAST summary of the Brundage Study is (emphasis added):

> In this study, bullets were fired from 10 consecutively manufactured 9 millimeter Ruger P-85 semi-automatic pistol barrels. Each of 30 examiners received a test set containing **20 questioned** bullets to compare to a set of **15 standards**, containing at least one bullet fired from each of the 10 guns. Of the **300 answers returned**, there were no incorrect assignments and one inconclusive examination.

This is not correct. The Brundage study consisted of 15 questioned bullets compared to a set of 10 standards (two test fired bullets from each standard set). This test was sent to 30 examiners and 450 answers returned (30 examiners x 15 questioned bullets) with no false positives and one inconclusive conclusion.

Hamby Study

The Hamby Study was a continuation of the Brundage study. Hamby et al. used the same firearm and ten consecutively manufactured barrels to produce an additional 240 test sets. The PCAST summary of this study states (emphasis added):

> In this study, bullets were fired from 10 consecutively rifled Ruger P-85 barrels. Each of **440 examiners** received a test set consisting of 15 questioned bullets and two known standards from each of the 10 guns. Of the **6600 answers returned, there were 6593** correct assignments, seven inconclusive examinations and no false positives.

This study combined the conclusions from the Brundage study, and additional results collected with both the original Brundage test sets and the 240 new test sets. If we subtract the original 30 responses from the Brundage study, the Hamby et al. article reports an additional 477 examiners having completed the test, for a total of 7155 answers with 7148 correct assignments and 7 inconclusive conclusions.

Fadul Pistol Slides Study

The PCAST summary of the Fadul Pistol Slides Study:

> In this study, bullets were fired from 10 consecutively manufactured semi- automatic 9mm Ruger pistol slides. Each of 217 examiners received a test set consisting of 15 questioned cartridge cases and two known cartridge cases from each of the 10 guns. Of the 3255 answers returned, there were 3239 correct assignments, 14 inconclusive examinations and two false positives.

This summary is correct; however, it is incomplete because it only includes Phase 1 of the study. It does not include the second phase of the study, the durability study. Results for Phase 1 and 2 are included in the same report. In Phase 2, an additional 114 examiners participated. The examiners received 5 more questioned cartridge cases (after the firearm had been fired 1000 times) and were asked to compare these cartridge cases to the 10 cartridge cases from the knowns that were previously received. A total of 570 answers were returned with 564 correct assignments, 5 inconclusive and one false positive.

Fadul EBIS Barrels Study

The PCAST summary of this study states (emphasis added):

> The 165 examiners in the study were asked to assign a collection of **15 questioned** samples, fired from **10 pistols**, to a collection of known standards; two of the 15 questioned samples came from a gun for which known standards were not provided.

This is not correct. Each test consisted of two known standards from each of the 8 pistols and 10 questioned samples. One of the known pistols had no matching questioned samples. Additionally, two of the unknowns had no matching known pistol.

Fadul et al. reported an overall error rate of 0.7% (95% lower bound 0.2%, 95% upper bound 1.2%).

From: [redacted]
To: [redacted]
Subject: Forensic Science
Date: Wednesday, December 14, 2016 6:10:22 PM
Attachments: 2016 Mnookin Error Rates for Latent Fingerprinting as a Function of Visual Complexity and Cognitive Difficulty.pdf
34 - TRIPLETT - AISOCC-JournalOfColdCaseReview-Volume2-Issue1-Jan2016.pdf

---

I'm sorry I have little time to reply due to a very sick relative but this is important so I'm sending a brief response.

PCASTs only error was in thinking fingerprints has been validated. The 2 studies referred to are only looking at conclusions, not how the conclusions were arrived at. They showed that conclusions are fairly reliable but more important is when conclusions are not reliable (for all pattern evidence disciplines). Conclusions are less reliable as the comparison becomes more complex (see Mnookin article attached). The next step is stating when comparisons are complex, which can easily be accomplished but is only being done in a small percentage of labs (see Triplett article attached).

Reliability is good, but knowing when and which conclusions are reliable is the only way to strengthen forensics.

Sincerely,
Michele Triplett
Forensic Operations Manager
King County AFIS Program

Complexity, Level of Association and Strength of Fingerprint Conclusions

By Michele Triplett [1]

Abstract

False convictions and false incarcerations have pushed the topic of forensic errors into the national spot light. Friction ridge comparisons (referred to as fingerprints for the remainder of this paper) are very accurate but errors have occurred. The strength of any conclusion needs to be indicated since criminal proceedings rely heavily on this type of information. The following paper discusses a possible explanation for errors and offers a more accurate and transparent approach for arriving at and reporting results. The proposed approach labels the complexity and demonstrable level of association found between two impressions which allow others to more accurately discern the strength of a conclusion.

Keywords: cold cases, fingerprint comparison, false convictions

[1] Michele Triplett is the Forensic Operation Manager for the King County Regional AFIS Program in Seattle, WA. She is a Certified Latent Print Examiner and holds a BS in Mathematics and Statistical Analysis. She has been employed in the friction ridge identification discipline since 1991 and is actively involved in several committees, organizations and educational events.

Standard Conclusions

Historically, fingerprint conclusions have been reported in a categorical fashion, such as 'the impression has been identified to John Doe'. Reporting conclusions in this manner has made conclusions sound conclusive, when in reality they may be strongly supported with visual data, marginally supported with visual data, or lack visual data that can be successfully demonstrated to others (i.e., simply the beliefs of the practitioners stating the conclusion). In order to determine the strength of the conclusion, the basis behind the conclusion needs to be assessed. Conclusions have been reported categorically as a means of simplifying a very intricate process that was based on a large number of non-quantifiable variables. No statistical model has been able to express the strength of conclusions despite on-going and previous efforts dating back to the late 1800s.

Criterion of Inclusions

Sufficiency to establish an identification is commonly based on either a practitioner's own tolerance level or non-validated administrative point standards set by an agency. Even without a validated sufficiency threshold, past conclusions have seemed fairly reliable; opposing conclusions and errors appeared virtually nonexistent. With the advent of the internet, information sharing has become easier and the variation in practitioners' conclusions has become increasingly more apparent, conclusions are not as definitive as once claimed (*Jackson v. Florida*, 2015; Stacey, 2005; Possley, 2015).

Evaluating Correctness of Conclusions

The lack of a clearly defined criterion for arriving at conclusions makes it difficult to evaluate practitioner's conclusions; without a standard, there is no means of judging correctness. This is extremely concerning when people's liberties and lives are on the line. Currently, the only way to assess a conclusion is to ask for another practitioner's opinion; which is mistakenly viewed as a measure of accuracy. Repeating a conclusion is simply measuring whether or not the conclusion is acceptable to another practitioner; it is not establishing absolute truth.

Establishing Error Rates

In the last decade, millions of dollars have been spent on error rate studies. These studies have assessed the accuracy of practitioner conclusions when comparing manufactured impressions to ground truth conclusions. The studies did not compare the error rates of different methods for arriving at conclusions. The studies indicate that the error rate is low but perhaps higher than previously assumed. Some studies assessed the repeatability of supporting data but they have not evaluated the acceptability of the support behind the conclusions (e.g., an accurate conclusion arrived at illogically would have been determined to be correct for the purposes of the research).

In casework, the ground truth is never known; casework conclusions are labeled as errors when others disagree with the conclusion. Since the research studies are assessing a measurement that does not apply to casework, the results of these studies may not accurately represent the error rate for casework. More importantly, the significant question to attorneys, judges, and the person identified as depositing a fingerprint at a crime scene is not how often experts make errors, rather which conclusions, and which methods, are at a higher risk of error? In order to reduce error rates and strengthen forensic conclusions, improved research would compare the error rates of different methods in order to show which technique produces the best results.

Paradigm Shift

The time has come where it is now essential to establish standards for arriving at conclusions and clear articulation of the strength of subsequent conclusions. Doing so will improve conclusions and give the ability to measure the correctness of conclusions. Instead of oversimplifying conclusions as categorical variables (identification or exclusion), it is more appropriate to present decisions on a continuum that expresses the complexity of a comparison (e.g., *Basic*, *Advanced*, *Complex*) and the demonstrable level of association (such as: overwhelming, marginal, or none). The complexity of a comparison is important because it determines the extent of testing required to ensure the interpretation and amount of data hold up under a critical review. The results of the testing establish the acceptable level of association, which indicates the strength of a conclusion (e.g., a complex comparison does not indicate that a conclusion is weak, it indicates that addition quality assurance measures are required to establish a strong conclusion). It is possible to assess the complexity of an impression in isolation of a comparison; however, the complexity may change during a comparison, making a pre-comparison assessment of an impression unnecessary.

Measuring information with words instead of numbers may seem unusual however; this is common in disciplines that are unquantifiable. For instance, hospitals rate the condition of patients on a wording scale (critical, severe, good, fair, etc.). The words chosen are not simply at the discretion of the doctor, there are criteria for each category so that every doctor rates patients the same. For bone fractures, doctors do not simply report that a leg is broken; they rate the severity of the fracture

in words (compound, hairline, etc.). Again, there are specific definitions for each description to ensure fractures are rated the same. Additionally, doctors use a 4 stage scale when reporting the severity of a cancer diagnosis; with specific definitions for each rating. The pattern evidence disciplines can and should follow suit and report more than a conclusion. Adding information that indicates the strength of the association found would benefit all interested parties.

Scientific Criterion: Data and Testing Over-Confidence

The primary question asked regarding fingerprint comparisons is how much information is enough to establish an identification. As stated above, the answer is not a quantifiable number, however, the accepted criterion used by other non-quantifiable comparative sciences (i.e., based on analytical reasoning) fits well within the realm of fingerprint comparisons. The criterion is to ensure conclusions have sufficient justification within established fundamental principles, to hold up against strong scrutiny. This is often times referred to as general consensus, although the term general consensus can be misconstrued as meaning that the majority of people would arrive at the same conclusion. General consensus is better defined as the conclusion has been debated until all doubt has been resolved. Resulting conclusions may be referred to as inferences that are supported by data. The strength of an inference is determined by assessing whether the support behind the inference satisfies any doubts presented by others.

Conclusion

Conclusions based on specific criterion and vetted against rigorous scrutiny will preempt errors and make conclusions more trustworthy than conclusions based on personal thresholds and confidence levels. Clear thresholds also make it possible to judge the acceptable level of association used to support a conclusion; which helps assess the risk of error for each conclusion (example to follow). Measuring acceptance or rejection based on a criterion is a far more informative approach than judging conclusions based on the beliefs of other individuals. Ultimately, utilizing the following method will provide stronger conclusions and allow others to assess the strength of conclusions.

Simplicity/Complexity Scale (*Basic, Advanced, Complex*)

The following rankings are intentionally minimized into three groups for simplicity. The number of rankings could be expanded but has been found to be unnecessary because the minor differences of opinion that may occur are insignificant to the end result. The criterion listed for each ranking are based on the prevailing views, i.e. tenprint comparisons are considered Basic, latent comparisons are considered Advanced, and comparison based on highly ambiguous or minimal data are considered Complex. Comparisons between listed rankings can be labeled as semi-advanced or semi-complex.

Those using this method must be trained in fingerprint comparisons in order to determine the region and orientation of impressions. Users must be trained in scientific protocols in order to understand concepts such as the amount of adequate testing required. For example, scientific conclusions are never based on one piece of data, such as excluding a person as the source based on the pattern type alone. Plausible conclusions must be tested before arriving at a well-supported conclusion. The testing required for each ranking is based on standard testing requirements for non-quantifiable comparative sciences (ensuring the conclusion holds up to rigorous scrutiny). Demonstrating the basis behind a conclusion is required upon any request.

The determination that the conditions for each ranking are met is not at the discretion of

the practitioner. Whether or not the conditions are met must hold up under rigorous scrutiny.

Many who attempt to rank comparisons in this manner quickly find that this is no different from how they have assessed images in the past. The main difference is that the weight of a conclusion is put in the data, not in the practitioner's beliefs or abilities, which protects against over interpretation and errors.

I. *If* There is sufficient data to establish, not presume, the region and orientation; and;

The data being interpreted consists of clear Galton points, spatial relationship, and intervening ridges; and

The correlation of data would easily be interpreted by others; and

The amount of information is large, not all data needs to be assessed or utilized (such as the majority of tenprint to tenprint comparisons)

Then **The comparison is considered *Basic***

Testing (such as consultation, corroboration, supporting documentation, or testing against strong scrutiny) is not scientifically required for this simplistic of conclusions, a practitioner can determine if the data used and the conclusion will meet the criteria (ID: Holds up to strong scrutiny, Exclusion: region, orientation and a clear target group of minutia).

A review of the conclusion is not necessary but may be set by agency policy.

Examples: Standard tenprint comparisons, comparisons with dissimilarities/discrepancies may be considered *Basic* when the area with the discrepancy is not needed to perform a comparison and arrive at a conclusion (the appearance of differences/discrepancies may exist but the reason unknown. Differences/discrepancies do not necessarily indicate a comparison overall is advanced, complex, or that an identification is not warranted).

Latent print comparisons where the region and orientation are known and the features are very clear and large (more data than necessary) are considered *Basic.*

II. *If* There is insufficient data to establish, not presume, the region and orientation (making the search more difficult); or

Ancillary features (scars, creases, incipient ridges) are being interpreted; or

The interpretation of data has slight ambiguity (may not initially be interpreted the same by others); however, the interpretation of data can easily be demonstrated to the satisfaction of others

Then **The comparison is considered *Advanced***

Testing (such as consultation, corroboration, supporting documentation, or testing against strong scrutiny) is optional but recommended (since the interpretation of data can easily be demonstrated); a practitioner can determine if the data used and the conclusion will meet the criteria (ID: Holds up to strong scrutiny, Inconclusive: No consistency found, Exclusion: region, orientation and a clear target group of minutia or multiple target areas if ambiguity is present).

A review of the conclusion is not necessary but may be set by agency policy to ensure appropriate testing.

Examples: Standard latent comparisons, known impressions deposited with extreme deposition pressure, twisting or smearing, complete tonal reversals, the use of creases, or relying on mostly ancillary features.

III. *If* The interpretation of data (Galton or ancillary features) has predominant ambiguity (the interpretation of data is questionable making it difficult to demonstrate to the satisfaction of others); or

The correlation of data is extremely limited (making it necessary to use rarity, ridge shapes, edges, pores, or features in simultaneous impressions)

Then **The comparison is considered *Complex***

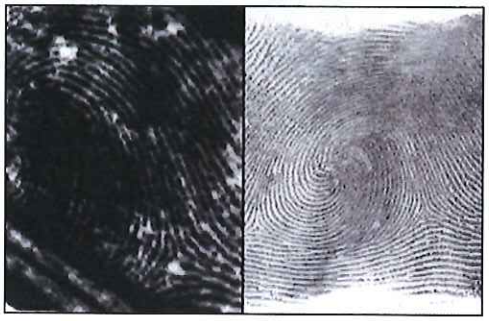
Testing (such as consultation, corroboration, supporting documentation, or testing against strong scrutiny to establish a consensus conclusion) is required to arrive at a conclusion that is well supported and tested under intense scrutiny.

A review is essential to ensure the appropriate amount of testing was performed.

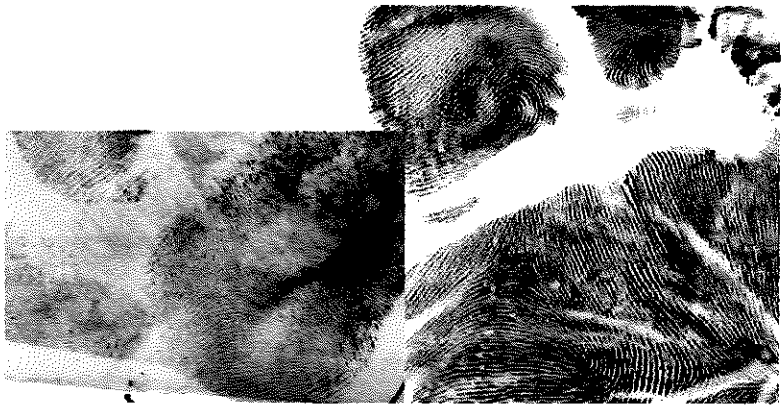
Examples: Tonal shifts, relying on highly ambiguous data (SCRO, Mayfield, Daoud). *Note*: Complexity is distinguished from difficulty in that difficulty level is based on a person's ability while complexity is based on the data in the impressions (either the unknown or known).

Below are examples of Basic Comparisons. The region and orientation are easily determined. The conclusion can be determined with the use of clear Galton points, their spatial relationship and the number of intervening ridges. The amount of information is abundant and not all data needs to be utilized (CLPEX.com fig's 36, 32, 40) (CLPEX, 2015). Conclusions from Basic Comparisons are very reliable.

The following example may be at the high end of Basic or Semi-Advanced. The region and orientation can be presumed. The conclusion can be determined with the use of Galton points, their spatial relationship and the number of intervening ridges. The amount of information is large and not all data needs to be assessed (FBI fingerprint image).



The comparison below may be considered Advanced since the region and orientation are not standard. However, the features within the image are clear and plentiful (CLPEX.com fig 68) (CLPEX, 2015).



The comparisons below fall into the category of Complex because the features within the unknown impressions are ambiguous, the interpretation of data may not be successfully demonstrated to others (CLPEX.com fig 95, Mayfield fingerprint comparison) (CLPEX, 2015; Saks & Koehler, 2005). Testing the interpretation of data for acceptability is essential to establish the appropriate conclusion.

The complexity of a comparison is based on the amount of ambiguity. The acceptable level of association is based on demonstrability and/or testing performed, which in turn determines the strength of a conclusion. The chart below can be used as a quick reference guide.

SIMPLICITY/COMPLEXITY SCALE: QUICK REFERENCE CHART

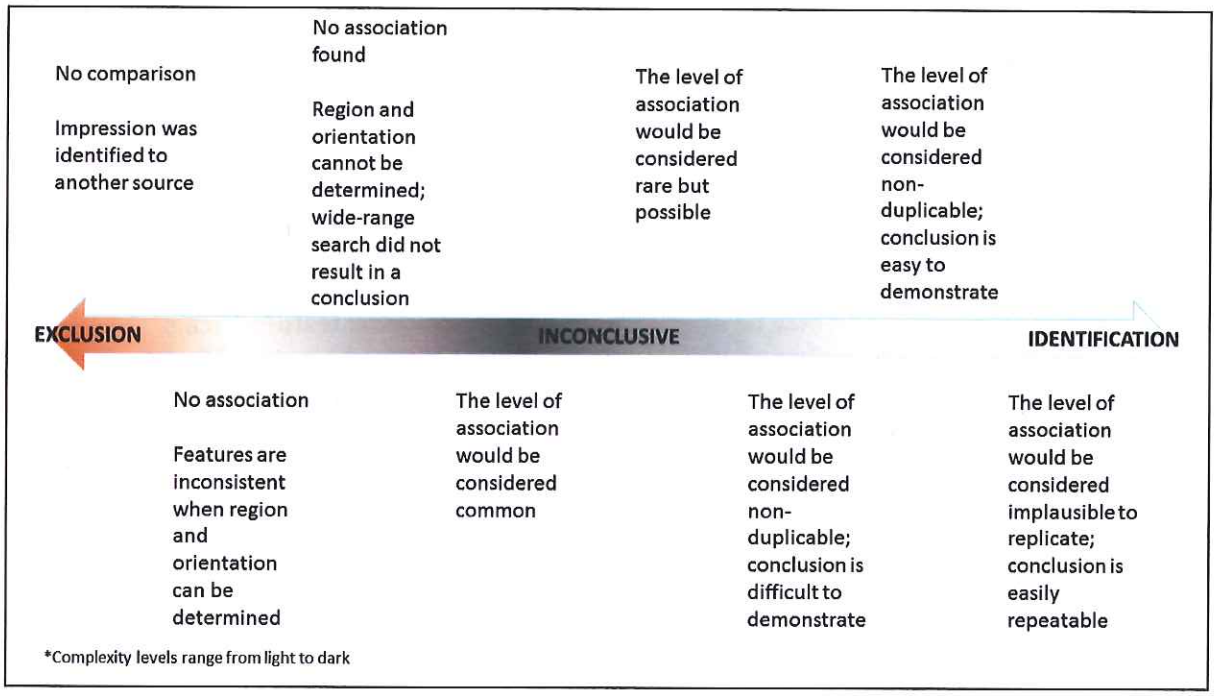
BASIC
Region and orientation can be determined. Use of Galton features, spatial relationship and intervening ridges (assessment of other features is not needed). Not all data needs to be assessed. Testing against scrutiny not required.

ADVANCED
Region and orientation may be questionable. Use of additional features (scars, creases, incipient ridges) or additional aspects (clarity, slight ambiguity of features). Not all data needs to be assessed. The use of ancillary features may be considered at the high end of Advanced. Testing against scrutiny recommended.

COMPLEX
Galton features are predominantly ambiguous. May include the use of edges, pores, or simultaneous impressions due to limited correlation of Galton features. Testing against scrutiny required.

Level of Association Continuum

It may seem reasonable to assume that erroneous identifications are more likely to occur as the level of association decreases (close non-matches; the gray ranges in the level of association continuum); however, this is not the case. Research into past identification errors demonstrates that misinterpretation of ambiguous data and reliance on reproducibility as the test for acceptability are the primary causes of errors. Past errors were found and acceptable associations established by testing the conclusion to ensure the interpretation of data holds up against strong scrutiny, ensuring the basis for the conclusion can be demonstrated to the satisfaction of others (i.e., general consensus) (Stacey, 2005; CBS Interactive, 2012).



Some agencies state the number of Galton points as an attempt at providing a weight to their conclusion. Stating a number of Galton points can be very misleading because it implies a weight that may not actually exist. A high correlation does not establish the strength of a conclusion because the assessment of those points may be a misinterpreted, as seen with the Mayfield error and the Dandridge error (Possley, 2015). The level of association is only meaningful if it can be successfully demonstrated to others, as required by the standard for non-quantifiable sciences.

Articulation of Conclusions

There is a lot of information that results from utilizing this method and therefore a variety of ways to articulate the information. For instance:

The comparison was considered:

a) basic
b) semi-advanced
c) advanced
d) semi-complex
e) complex

Testing performed:

a) none
b) tested against strong scrutiny for acceptable interpretation of data

The level of association is:

a) impression associated to another person, exclusion to this subject by deduction

b) overwhelming inconsistency, exclusion to this subject

c) features too broad to determine specific search area, no consistency found after a wide-range search

d) the level of association is limited or marginal, an amount of consistency seen in others

e) the level of association is high or considerable, not expected in others but plausible (may be referred to as an investigative lead or a person of interest)

f) the level of association is persuasive, difficult to demonstrate but considered implausible to replicate

g) the level of association is compelling, easy to demonstrable, and considered implausible to replicate

h) the level of association is overwhelming, easily repeatable by other experts, and considered implausible to replicate

Specific conclusion in casework could be articulated as one of the following:

"The comparison is *Basic.* The level of association is overwhelming and easily repeatable by others."

Or, "The comparison is *Advanced.* The level of association is *compelling,* easy to demonstrate, and considered implausible to replicate."

Or, "The comparison is *Complex. Testing against strong scrutiny* determined the level of association to be persuasive and considered implausible to replicate."

Conclusions presented with this type of information demonstrate to others that the practitioner relied on criteria and demonstrable data to protect against over-interpretation and to ensure conclusions are as solid as humanly possible. This method can also be beneficial to re-assess conclusions arrived at using a different method. The level of complexity, the degree of testing performed, and the level of association will establish the strength behind any conclusion.

The well-known 2004 FBI erroneous identification to Brandon Mayfield can be assessed under this method. Under this method, the identification to Mayfield would have been labeled complex since many of the associations used were ambiguous. A complex rating indicates that testing against rigorous scrutiny is essential. Rigorous scrutiny was not performed by the FBI since the culture at that time discouraged disagreement among examiners (Stacey, 2005). The Spanish experts ap-

proached their review in a more critical fashion by questioning the interpretation of data and the conclusion. If rigorous testing against scrutiny had occurred within the FBI or by the external practitioner reviewing the comparison for Mayfield, then the conclusion would have been labeled 'e' at best. If it had been known that it was a complex comparison and rigorous testing against scrutiny had not been performed, yet a conclusion of 'f', 'g' or 'h' was being reported, then others would clearly see the red flags in this case. Other past errors can be tested against this system as well. Each would show that an identification would not have held up under this standard.

References

CBS Interactive. (2012, November 4). Lana Canen freed over bad fingerprint evidence after 8 years in prison for Indiana murder. *CBS News.* Retrieved from http://cbsn.ws/1MItPhY

CLPEX. (2015). *Complete latent print examination.* Retrieved from http://www.clpex.com.

Federal Bureau of Investigation. (2013, July 9). *Latent hit of the year award: Fingerprint tool helps solve 1999 murder.* Retrieved from http://1.usa.gov/1WYhuaD

Kim Jackson v. State of Florida, SC13-2090 (Fla. 2015). Retrieved from http://fla.st/1H1VJVK.

Possley, M. (2015, October 12). Beniah Alton Dandridge. *The National Registry of Exonerations.* Retrieved from http://bit.ly/1Mor8AH.

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science, 309*(5736). DOI: 10.1126/science.1111565

Stacey, R. B. (2005). Report on the erroneous fingerprint individualization in the Madrid train bombing case. *FBI Forensic Science Communications.* Retrieved from http://1.usa.gov/1LfhKv3.

Correspondence concerning this article should be addressed to:

Michele Triplett, 516 3rd Ave, Mail Stop: KCC-SO-0100, Seattle, WA 98104, Email: s.triplett@comcast.net

How to cite this article:

Triplett, M. (2016). Complexity, level of association and strength of fingerprint conclusions. *Journal of Cold Case Review, 2*(1), 6-15.

ERROR RATES FOR LATENT FINGERPRINTING AS A FUNCTION OF VISUAL COMPLEXITY AND COGNITIVE DIFFICULTY

NIJ Award 2009-DN-BX-K225

Jennifer Mnookin (P.I.), Philip J. Kellman (Co-P.I.), Itiel Dror, (Co-P.I.)
Gennady Erlikhman, Patrick Garrigan, Tandra Ghose, Everett Metler, Dave Charlton

Abstract

Comparison of forensic fingerprint images for purposes of identification is a complex task that, despite advances in image processing, still requires highly trained human examiners to achieve adequate levels of performance. Latent fingerprints collected from crime scenes are often noisy, distorted, and represent only a portion of the total fingerprint area, making matching tasks difficult. While it is clear that expertise in fingerprint comparison, like other forms of perceptual expertise, such as face recognition or aircraft identification, depends on perceptual learning processes that lead to discovery of features and relations that matter in comparing prints, relatively little is known about the perceptual processes involved in making fingerprint comparisons, and even less is known about how the visual characteristics of fingerprint pairs relate to comparison difficulty. This project aims to determine more about the relationship between the measurable, visual dimensions of fingerprint pairs and the level of comparison difficulty for human examiners, both experts, and to a lesser degree, novices. For this research, we assembled a new database of latent fingerprints, matching tenprints, and close, non-matching tenprints. Using this database, we measured expert examiner performance and judgments of difficulty and confidence in a variety of settings. For the experts, we developed a number of quantitative measures of image characteristics and used multiple regression techniques to discover predictors of error as well as perceived difficulty and confidence. A number of useful predictors emerged, including variables related to image quality metrics, such as intensity and contrast information, as well as measures of information quantity, such as the total fingerprint area. Also included were configural features that fingerprint experts have noted, such as the presence and clarity of global features and fingerprint ridges. Within the constraints of the overall low error rates of experts, a regression model incorporating the derived predictors demonstrated reasonable success in predicting difficulty for print pairs, as shown both in goodness of fit measures to the original data set and in a cross validation test. The results indicate the plausibility of using objective fingerprint image metrics to predict expert performance and subjective assessment of difficulty in fingerprint comparisons. We also examined the extension of these results to settings that better approximate real-world fingerprint examiner scenarios, and found our regression model continued to provide significant explanatory value for a substantial portion of the prints. While further research is necessary, this research provides strong support for the plausible but previously untested assumption that for expert fingerprint analysis, difficulty (and by extension, error rate) is in significant part a function of measurable, visual dimensions of print comparison pairs. In addition to this primary focus, we also conducted several extensions to this research, involving expert metacognition and novice comparison. These experiments showed that experts have substantial, albeit imperfect, subjective knowledge about the difficulty of print pairs. Our experiments also showed that novices perform very poorly and showed no consistent pattern of feature use. This research thus also contributes to our understanding of the source and extent of human expertise in latent fingerprint analysis.

Table of Contents

Executive Summary

There has been a longstanding belief in the scientific validity of fingerprint evidence, based both on the apparent permanence and uniqueness of individual fingerprints and on the experience-based claims of trained fingerprint examiners. In the past, fingerprint evidence, in the hands of an experienced examiner appropriately applying the methods of the field, was often claimed to be "infallible" or to have a "zero error rate" (Cole, 2005; Mnookin 2008b). Yet systematic scientific study of the accuracy of fingerprint evidence is a rather late development, still very much in progress. The traditional claim of infallibility for fingerprint identification has been brought to the spotlight and questioned in light of high-profile cases in which errors have been discovered. While it is likely that well-trained, experienced examiners are highly accurate when making positive identifications, it is also clear that errors still occur. Recently, with the National Academy of Sciences (2009) inquiry into forensic science, new research has begun to emerge. The available data now suggest a low level of erroneous match determination by experts under experimental conditions and a higher rate for erroneous exclusion determinations (e.g., Ulery, Hicklin, Buscaglia, & Roberts, 2011; Tangen, Thompson, & McCarthy, 2011).

At present, however, fingerprint identification is a strikingly subjective process (e.g., NIST 2012, chapter 3). There are not validated, objective metrics for any meaningful step in the comparison process, from the determination of whether there is sufficient information to warrant a comparison to the final judgment of match or non-match (e.g.,Mnookin, 2010). Not only are these judgments and conclusions subjective, but at present, there is no method by which the relative simplicity or difficulty of a print pair can be determined. While examiners may, based on their experience, have a view about the relative ease or difficulty of a comparison, whether these subjective judgments are warranted has not been known. The main goal of the project was therefore to identify and quantify fingerprint image features that are predictive of identification difficulty and accuracy. A quantitative way of assessing fingerprint image quality and comparison difficulty would be an extremely useful development. First and foremost, objective metrics for measuring difficulty create the possibility of associating error rates with the level of difficulty. It is a matter of common sense to recognize that if some print comparisons are unusually hard, examiners will therefore be more prone to make possible mistakes in their analysis; conversely, easy comparisons would be expected to produce fewer errors than hard ones. But common sense or not, prior to this research, no research focused on examining this issue, or attempting to look at the relationship between error rate and difficulty.This project provides foundational steps toward the possibility of associating error rates with difficulty. An objective metric of difficulty has other benefits as well. Such a metric can be used to alert examiners when additional care is warranted (i.e., for a particularly difficult comparison), to caution examiners who are inclined to label a print pair as inconclusive that further examination might be prudent, and to create a set of fingerprint images with objective difficulty ratings that can be used for training examiners.

In addition to creating a fingerprint image quality metric, the project had several other objectives: (1) to create a realistic fingerprint image database with known ground truth (i.e., true matching prints and close non-matches); (2) to add to the rather modest scientific literature investigating fingerprint expertise; that is, to assess objectively expert performance both in terms of accuracy in fingerprint identification and in terms of the image characteristics that related to performance; (3) to examine the relationship between objective expert performance and subjective assessments of difficulty and confidence, to evaluate experts' metacognitive abilities vis-à-vis the comparisons they make; (4) to compare the use of visual information by experts in their fingerprint analysis to the use made by novices.

The report is roughly divided into two sections. In the first part, we identify and describe the computations for several image features that we hypothesized might correlate with identification accuracy. The features were a mixture of image properties such as contrast and brightness levels, traditional (i.e., Level I, II, and III) fingerprint features such as visibility of deltas and clarity of ridges, and relational features that applied to the *pair* of prints in a comparison. The final point is important since a comparison does not depend only on the quality of the latent, but also on the shared information between it and the known print.

The fingerprint images that were used were latent prints collected from volunteers. Each volunteer left several impressions on a variety of glass objects. Practicing fingerprint examiners verified that the impressions we collected were realistic exemplars of the kinds of images that could be found in forensic settings. Corresponding known prints were also collected from each individual, providing a set of known matching prints. The latent prints were submitted to an AFIS system and close, non-matching prints were selected for each of the collected latents. The final print database contained over 500 pairs of matching and 500 pairs of non-matching prints.

In the second section, we describe three behavioral studies in which we measured expert examiner and novice performance on a subset of the database. In Experiment 1, expert examiners made timed comparison judgments for print pairs via an online interface that we designed for this purpose. Examiners were shown a pair of prints and made an identification or exclusion judgment and provided difficulty and confidence ratings for each such comparison. We collected data from 56 examiners at a forensic conference. Examiner accuracy was high: there were 200 errors made out of 2292 total comparisons (overall accuracy of 91%) with more false negative or incorrect exclusions (14%) than false positives or incorrect identifications (3.2%). Difficulty and confidence ratings were highly correlated with accuracy, indicating that experts were often able to identify which comparisons were difficult or likely to be error-prone. We fit a regression model to the accuracy data to predict examiner performance from image features. We identified several features that were predictive of accuracy, including the ratio of the image areas of the latent and known print, the combined reliability of ridge information in both images, the variability of contrast across small regions of both images, and the visibility of deltas in the latent print. The model was also fairly successful in predicting the accuracy on a held-out set of fingerprints that were not used in the regression ($R^2_{adj} = 0.64$). A separate, classification analysis was able to identify print pairs for which at least one examiner made a mistake with a 75% (15/20) accuracy.

Experiment 2 sought to compare expert performance with that of novices. Naïve participants with no fingerprint examination training made fingerprint comparisons in an interface we designed. Not surprisingly, they performed far below experts, indeed approximately at chance. For a second group of subjects, we showed a brief video that highlighted various fingerprint features that could be used in making comparisons such as *minutiae*. Participants who viewed the video only showed a mild improvement in performance. However, there were noticeable changes in their biases to make an identification; in particular, trained participants seemed more hesitant to label a fingerprint pair as coming from the same source, perhaps because the training video emphasized the cost of making an incorrect identification and made the difficulty of fingerprint examination more apparent. We fit the novice data to the same regression model from Experiment 1. We found that untrained novices relied on different features than experts and that some features that experts used had the opposite effect on novice performance. Trained novices had the weight given to certain features shifted closer to those of experts, indicating that with training novices may learn which image information is relevant for identification, while learning to ignore irrelevant features.

Experiment 3 served as an extension and validation of Experiment 1. We used the data from Experiment 1 to make performance predictions for a new subset of fingerprints from the database.

We also designed a new web interface that incorporated many of the image processing tools that are typically available to experts. For example, website users could reverse the contrast of the images, zoom in and out, mark *minutiae*, rotate the image, and increase or decrease brightness and contrast. In addition, we gave examiners unlimited time for each comparison. We recruited a new set of 34 expert examiners and allowed them to perform the experiment on their own time. Performance was slightly higher than in Experiment 1, with 10% incorrect exclusions and 0.25% incorrect identifications. Fitting a regression model identified many of the same predictors as in Experiment 1. Using the model from Experiment 1 to predict performance for this set of prints was qualitatively successful for comparisons that were not labeled inconclusive by any examiners, but performed poorly for those that were. Further work needs to be done to incorporate image processing tool use into the regression model and to account for performance on comparisons that are labeled inconclusive.

We have taken several foundational steps in this project. First, and most importantly, we have successfully demonstrated that there are image features that can be quantified and used to predict examiner performance. While we do not claim our list to be complete or exhaustive, this research is a vital first step in proving that such features can be found and provides a useful guide for future research. In addition, these experiments provide persuasive evidence that there is a meaningful and important correlation between comparison difficulty and error rate. Errors were not distributed randomly across our exemplars, but rather, significantly clustered, revealing that difficulty is, to a significant extent, a function of visual aspects of the specific comparison. This recognition also provides evidence that it is not especially helpful to seek a field-wide "error rate" for latent fingerprint identification. Instead, as our scientific knowledge of fingerprint identification continues to progress, it will be more useful to seek error rates for different categories of comparisons, based on objective difficulty level.

Second, we have provided useful evidence for and quantification of examiner expertise, both in controlled (Experiment 1) and more realistic (Experiment 3) settings. This has also allowed us to examine inter-rater reliability and metacognitive judgments (i.e., whether examiners are aware which prints are actually difficult) and to contrast performance with novices (Experiment 2). Furthermore, similarity in examiner performance between Experiments 1 and 3 suggests that one can study examiner expertise without needing to perfectly replicate work conditions and still get a useful estimate of performance. This concern has previously limited research, and the comparative experiments in this study provide valuable information for future researchers to consider when engaging in study design. Third, we found preliminary evidence that even a small amount of training can adjust novice performance. This suggests that with more extensive bursts of training we can examine, in a controlled manner, the process by which an examiner becomes an expert. This will allow for various interventions in the training process, allowing for more efficient and automated training techniques. Fourth, we have constructed an independent fingerprint database, with ground-truth and difficulty ratings that can be used for future studies or for examiner training. In addition, we have created several web-deliverable tools that can also be used for evaluation or training that replicate many of the image processing features available to examiners.

Beyond the scientific findings, there are important aspects of the research that can impact policy. A more sophisticated understanding of the relationship between error rate and difficulty is, or should be, important for the courts in weighing fingerprint evidence. Courts are instructed, when assessing expert evidence, to focus on the "task at hand", and this research helps to show that fingerprint examination may vary in difficulty in ways that may be relevant to its evaluation as evidence (Daubert vs. Merrell Dow Pharmaceuticals, 1993; Kumho Tire Co. vs. Carmichael, 1999). More nuanced assessments of fingerprint task difficulty might, for example, affect how a judge understands admissibility of that specific conclusion, or what degree of certainty the expert will be

allowed to express, or it might appropriately impact the weight given to a specific match conclusion by the fact-finder (Faigman, Blumenthal, Cheng, Mnookin, Murphy & Sanders, 2012).

The implications of these findings go beyond the court; they provide vital insights that can considerably enhance the procedures used in forensic laboratories. For example, similar to medical triage, the need for different procedures and checks can be made to fit the difficulty of the comparison. The understanding of what makes some comparisons more difficult than others also has implications for the selection and training of fingerprint examiners. During selection, benchmarks and skill sets can be set as criteria to ensure candidates have the acquired the necessary cognitive abilities needed to perform their job adequately. In addition, in evaluating the significance of errors for trainees, better information about difficulty level will be of great assistance. Trainees who make mistakes on simpler stimuli can be distinguished from those whose errors occur only on more difficult materials; for evaluating performance, all errors are not – and should not be treated as – equal.

While further research is clearly necessary to build on these results, this research provides significant steps forward for helping to establish that error rates are related to difficulty; for beginning to provide evidence for what visual dimensions of fingerprint comparison pairs are associated with difficulty; and for helping to tease out both examiner's metacognitive abilities and the substantial degree of examiner expertise in this domain.

I. Introduction

There has been a longstanding belief in the scientific validity of fingerprint evidence, based both on the apparent permanence and uniqueness of individual fingerprints and on the experience-based claims of trained fingerprint examiners. In the past, fingerprint evidence, in the hands of an experienced examiner appropriately applying the methods of the field, was often claimed to be "infallible" or to have a "zero error rate" (Cole, 2005; Mnookin 2008b). Yet systematic scientific study of the accuracy of fingerprint evidence is a rather late development, still very much in progress. The traditional claim of infallibility for fingerprint identification has been brought to the spotlight and questioned in light of high-profile cases in which errors have been discovered. While it is likely that well-trained, experienced examiners are highly accurate when making positive identifications, it is also clear that errors still occur. Recently, with the National Academy of Sciences (2009) inquiry into forensic science, new research has begun to emerge. The available data now suggest a low level of erroneous match determination by experts under experimental conditions and a higher rate for erroneous exclusion determinations (e.g., Ulery, Hicklin, Buscaglia, & Roberts, 2011; Tangen, Thompson, & McCarthy, 2011).

Contrary to popular belief and its depiction on many television shows, fingerprint identification – matching a fingerprint from a crime scene to one on file – is not a fully automated process. While algorithms can compare known prints (fingerprints collected in controlled conditions such as in a police station where the fingerprint images are clear) with high accuracy, identifying latent prints (those found at a crime scene) falls to individual fingerprint examiners who are extensively trained (Vokey, Tangen, & Cole, 2009). However, the nature and extent of examiner expertise has only recently come under scientific scrutiny (e.g., Busey & Parada, 2010; Busey & Vanderkolk, 2005; Dror & Charlton, 2006; Dror, Charlton, & Péron, 2006; Dror, Champod, Langenburg, Charlton, Hunt, & Rosenthal, 2011; Tangen, Thompson, & McCarthy, 2011; Ulery, Hicklin, Buscaglia, & Roberts, 2011). While several proficiency tests have been used to evaluate expertise, many may have used an overly limited number of prints; this may have led to inaccurate estimates of examiner performance because of idiosyncratic fingerprint properties that made a particular identification easy or difficult (Cole, 2006, 2008; Vokey, Tangen, & Cole, 2009).

Mistakes in fingerprint matching are costly and can put lives and livelihoods at risk. Errors in fingerprint matching are of two types that have different implications. A *false negative*, where a matching pair is labeled as non-matching, could, in a criminal proceeding, allow a guilty suspect to be set free. A *false positive*, where a non-matching pair is labeled as a match, could lead to, in a criminal proceeding, the conviction of an innocent person. Existing data suggest that fingerprint experts err more on the side of false negatives (about 8% of total judgments made on a match/non-match task for fingerprint pairs) than false positives (about 0.1%) (Langenberg, 2009; Tangen, Thompson, & McCarthy, 2011; Ulery, Hicklin, Buscaglia, & Roberts, 2011). Experts perhaps tend to incorporate the presumption of innocence, erring on the side that would free the guilty rather than convict the innocent, although false positive rates are not zero.

The practical importance of understanding when and why fingerprint comparison errors occur is likely to increase as technology advances. Current Automated Fingerprint Identification Systems (AFISs) retrieve from a database a number of prints associated with known individuals that could be potential matches for a particular latent. Under typical procedures, the intervention of a human expert is required for deciding which if any candidates generated by an AFIS comprise a match to a latent. Candidate matches selected by a properly functioning AFIS should often appear similar to the latent entered into the system, a fact that likely increases the potential for human error. Imagine, by way of contrast, a situation in which an examiner is asked to compare a latent print to a known print

of a particular suspect in a criminal case. Assuming the two prints are not from the same individual, it would be a remarkable coincidence if the prints were highly similar. Use of an AFIS has high value in extracting candidates from a database, but it puts the examiner in the position of routinely needing to distinguish actual matches from close (highly similar) non-matches. The likelihood of human error increases with the degree of similarity of the potential candidates extracted by AFIS, thereby making the comparison process more difficult (Ashworth & Dror, 2000; Vokey, Tangen, & Cole, 2009). With the increase in size of AFIS databases, the possibility of finding a look-alike non-match increases, thereby increasing the potential for false positive errors (Cole, 2005; Dror & Mnookin, 2010; Dror, Péron, Hind, & Charlton, 2005).

From a visual information processing perspective, it is interesting and important to determine what visual characteristics of fingerprints influence the ease and accuracy of comparisons. Ultimately, it may be possible to evaluate a fingerprint comparison in terms of the quantity and quality of visual information available (Pulsifer et al., 2013) in order to predict likely error rates and to assess when there is insufficient information to warrant any conclusion.

Perceptual Aspects of Fingerprint Expertise

If asked to give reasons for a conclusion in a given comparison, fingerprint examiners will report significant explicit knowledge relating to certain image features, such as global configurations, ridge patterns and minutiae, as these are often explicitly tagged in comparison procedures. They are also pointed out during training of examiners. It would be a mistake, however, to infer that the processes of pattern comparison and the determinants of difficulty are in general available for conscious report or explicit description. As in many other complex tasks in which learning has led to generative pattern recognition (the ability to find relevant structure in new instances) and accurate classification, much of the relevant processing is likely to be at least partly implicit (Chase & Simon, 1975; Schneider & Shiffrin 1997; for a review, see Kellman & Garrigan, 2009).

Like many other tasks in which humans, with practice and experience, attain high levels of expertise, feature extraction and pattern classification in fingerprint examination involves *perceptual learning* -- experience-induced changes in the way perceivers pick up information (Gibson, 1969; Kellman, 2002). With extended practice, observers undergo task-specific changes in the information selected – coming to discover new features and relationships that facilitate classification in that domain. Evidence supporting this claim comes from increased perceptual learning when these features are exaggerated during training (Dror, Stevenage, & Ashworth, 2008). While several studies have explored the influence of bias and emotional context on fingerprint matching and classification (e.g., Dror, Péron, Hind, & Charlton, 2005; Dror & Charlton, 2006; Dror, Charlton, & Péron, 2006; Dror & Cole, 2010; Dror & Rosenthal, 2008; Hall & Player, 2008), there has been relatively little work investigating perceptual aspects of expertise among examiners or perceptual learning processes that lead to expertise.

There are also profound changes in *fluency*: What initially requires effort, sustained attention, and high cognitive load comes to be done faster, with substantial parallel processing and reduced cognitive load (Kellman & Garrigan, 2009). In turn, becoming more automatic at extracting basic information frees up resources for observers to discover even more subtle or complex structural information (see, e.g., Bryan & Harter, 1899). This iterative cycle of discovery and automaticity followed by higher-level discovery is believed to play a significant role in attaining the impressive levels of performance humans can attain in areas such as chess, chemistry, mathematics, and air traffic control, to name just a few domains (Kellman & Garrigan, 2009; Kellman & Massey, 2013).

These considerations motivate the research presented here. The primary goals were to: (1) create a fingerprint database with ground-truth (true matches) information and sufficiently difficult comparison to use as a testing base for future experiments that evaluate expert performance, (2) measure expert examiner performance on a variety of prints including difficulty comparisons, (3) measure novice performance to create a basis of comparison for expert skill, and (4) create a predictive framework by which one could assign an appropriate level of confidence in expert decisions derived from an objective assessment of characteristics of the pair of images involved in a particular fingerprint comparison. These goals are interconnected. Examiner performance levels (error rates) are likely to depend on the complexity and difficulty of the comparison: as comparisons get more difficult, errors are more likely to occur. Hence, the characterization and prediction of error rates should relate to the perceived difficulty of the comparison. Notwithstanding this relationship, no previous research on fingerprint identification has attempted to generate objective models for the assessment of perceived fingerprint comparison difficulty. Note that we use the term *comparison* difficulty advisedly. One of the insights guiding our research was that the right question is not merely whether a particular print is 'easy' or 'difficult,' clear or unclear, rich in information or less so. Rather, the right question is the difficulty of a given comparison. While latents may vary in quality more than tenprints, and thus may be the primary driver of difficulty, the specific comparison will also be relevant to determining difficulty. (To see the point most clearly, consider: a low quality print might nonetheless be part of an easy comparison when the tenprint is of a different pattern type. Similarly, a high quality latent might be part of a difficulty comparison when it bears an unusually high degree of similarity to the tenprint to which it is being compared.)

Several studies have attempted to quantify expert performance. Tangen, Thompson, and McCarthy (2011) generated a testing set of 36 simulated latent prints from the Forensic Informatics Biometric Repository. Twelve were paired with a corresponding known print match, 12 were paired with a randomly selected, non-matching print from the same database, and 12 were paired with similar prints found by submitting the latent prints to the Australian National AFIS. This resulted in a testing set in which the ground truth was known, i.e., for each latent print there was a corresponding, correctly matching known print. Thirty-seven experts and 37 novices made similarity ratings on a scale of 1 (different) to 12 (same). Judgments of "inconclusive" were not allowed. Only accuracy information was computed from the rating scale. Performance in the dissimilar and similar non-matching conditions was highest for experts, at 100% and 99.32% respectively. Performance was lower when the latent and known prints matched: 92.12%, indicating that experts were more likely to "free the guilty" than "convict the innocent", although both kinds of errors were made. Novice performance was markedly lower than experts'. Their accuracy was best in the match and dissimilar conditions, with accuracies of 74.55% and 77.03% respectively, and worst in the similar non-match condition, with an accuracy of 44.82%. Similar performance in experts between similar and dissimilar non-matches may reflect the results of training that is absent in novices.

Despite the interesting findings, and the large quantity of test images and examiners, several important questions remain unaddressed by the Tangen et al (2011) study. While the fingerprints were collected in a realistic manner by having individuals grasp objects, it is unknown whether the set of prints is sufficiently representative. As with proficiency tests, perhaps this set of prints was particularly easy or difficult if they did not, for example, capture a sufficient variety of smudges and distortions that might occur. The prints were generated by having individuals grasp objects or push open a door; these kinds of manipulations may have yielded a disproportionate amount of relatively clean fingerprints with little distortion. An expert (one of the authors) determined that all of the prints used in the study had sufficient information to make a judgment (i.e., would not be judged as "inconclusive"), but (through no fault of the authors, given the lack of objective metrics available) there was not any other way to determine difficulty. Importantly, one would like to be able to somehow assess the difficulty of fingerprint comparisons, to be able to determine when a

comparison should be easy and when it should be difficult and could lead to an increased error rate. For example, measuring novice performance only on matches and dissimilar non-matches would lead one to incorrectly estimate their average accuracy at comparing prints to be approximately 75%. The similar non-match condition in which accuracy is at chance is critical in demonstrating the difference between experts and novices. Without being able to quantify the degree of dissimilarity (difficulty of comparison) in the similar non-match condition, one can only say that expert performance is near perfect for this particular set of comparisons.

Other studies, using different fingerprint databases, have found novice accuracy to be closer to 85% (Vokey, Tangen, & Cole, 2009, Experiment 2). Discrepancy in accuracy estimates could be due to variability in the kinds of prints used for the study or the kinds of image manipulation tools (e.g., rotation of one of the images) available to participants. Without a quantitative measure of the properties of a fingerprint image that make a comparison difficult or easy, comparing accuracy rates across heterogeneous databases would provide little information about true ability, since the prints used could be substantially varied in difficulty.[1]

Such considerations naturally lead to the question of what features of the images make a particular comparison difficult or easy. For example, if many experts made errors in the match condition on the same fingerprints, it would be useful to know what features of those fingerprint images led to the errors. Identifying objective image features that correlate with accuracy would allow for predictions of comparison difficulty and could be used to tag print pairs that require additional scrutiny because they are more likely to be erroneously classified.

Ulery, Hicklin, Buscaglia, and Roberts (2011) have made an important first step toward addressing these issues. They created a large dataset of 744 print pairs including subjectively rated "low quality" latents that were rated as representative of those encountered in regular casework. In addition, the overall difficulty of comparisons was rated to be similar to casework by a majority of participants. A slightly greater proportion of images used in the study was rated as poor quality according to the NIST Fingerprint Image Quality Metric (NFIQ) compared to examples from AFIS. Non-matching pairs were selected by submitting latent pairs to an Integrated AFIS. 169 examiners participated, each evaluating approximately 100 randomly selected print pairs. Because the testing sets were generated randomly, there was large variability in the number of examiners that evaluated each pair. Examiners were given the option to label a comparison as "inconclusive". Among 4,985 non-match trials, there were 6 false positives (accuracy: 99.89%), each on a different comparison, made by 5 unique examiners. There were 611 misses (matches evaluated as non-matching) out of 8,189 comparisons (accuracy: 92.54%). These results were very similar to identification accuracy amongst experts in Tangen, Thompson, and McCarthy, (2009). Performance correlated with years of experience and certification suggesting that some variability is due to individual differences among experts (Ulery et al., 2011). Participants were also asked whether there was enough information in each latent image to make an identification, to make an exclusion (less information may be needed for exclusions since only one non-matching feature is needed between a latent and known print), whether an identification is possible conditional on the content of the known print, or whether there was not enough information in the latent to make a comparison (in which case the print was not shown in a comparison). For matching pairs, only 48% of latents were unanimously agreed to contain enough information to make an identification; agreement was 33% for non-matching pairs. One source of variability in performance is therefore individual differences among expert examiners. Some of this variability may be due to different amounts of expertise, since duration and type of

[1] Of course, the visual qualities that make comparisons easier or more difficult for novices may or may not bear much resemblance to the characteristics that make prints difficult for experts. One of the experiments discussed below (Experiment 2) has relevance to this point.

training correlate with performance. Other differences may be due to lack of a standard for determining what counts as sufficient information. Without some way of measuring information content and quality, it is impossible to know what makes a comparison difficult, which comparisons actually are difficult (without relying on subjective ratings), and whether an examiner is correct in determining that there is sufficient or insufficient information to make a comparison. Similarly, Langenberg (2009) had six examiners complete 120 comparisons in two phases. He investigated overall accuracy, verification accuracy, consistency within and across examiners, as well as type of conclusion (identification, exclusion, inconclusive, or no value). The resulting performance data are interesting and informative. However, this study did not quantify what *features* of the images may have resulted in errors or disagreement among examiners. While it is important to know what the average accuracy of an average examiner may be on an average fingerprint, that was not a primary goal of our project. Rather, our effort was to identify, using objective measures, whether a particular comparison is easy or difficult and whether it is likely or not to result in an error.

A recent NIJ report has made a valuable early attempt at measuring fingerprint quality and information content (Neumann, Champod, Yoo, Genessay, & Langenburg, 2013). Almost 150 examiners evaluated 15 fingerprint pairs for information "sufficiency". Examiners who participated in the study were asked to classify, using a web-based tool, regions of the images that had low, medium, or high quality. They were also asked to mark, by hand, as many minutiae as they could find and to classify them. In addition, they were asked to make several subjective assessments of quality regarding fingerprint properties such as amount of distortion or degradation. The authors examined relationships between marked features (minutiae), perceived quality metrics, and the conclusions reached by examiners. Interestingly, there was a great deal of variation across examiners in terms of assessment of finger ridge quality, degradation and distortion, and the number of minutiae. The researchers were unable to find a quantitative measure common to all examiners that predicted whether there was sufficient information to reach a conclusion. Other features, including demographic factors, seemed to have little effect. This report underscores the problem that the features examiners selected were ultimately subjective, and therefore dependent on the idiosyncrasies of the specific examiners making the comparisons. That is, different examiners would produce different features for the exact same fingerprint image. This research, while interested in questions related to ours, highlights the importance of our project, in which we strove to identify objective (i.e., observer-independent) image features that were predictive of accuracy. The features we identified can be computed automatically for any fingerprint pair and involve neither a laborious and subjective period of minutiae marking and classification, nor the concerns that arise from any subjective process about inter-examiner consistency and reliability.

Fingerprint Features in the Standard Taxonomy

The first step in latent print examination is often manual preprocessing. For example, the region of the image that contains the fingerprint could be selected from the background and oriented upright. If a fingerprint is to be submitted to a database for automated comparison, key features need to be identified and labeled. Automated searches are then carried out by software that finds fingerprints on file with similar spatial relationships among labeled features in the submitted fingerprint. This is the only part of the examination and comparison process that is automated. The software returns a list of potential matches, some of which will likely be quickly excluded. Some will likely be closer non-matches or a match, and these require further scrutiny by an examiner.

Whether examiners are provided with potential matches via automated database searches or via investigative work, they often make their match decisions using the approach: Analysis, Comparison,

Evaluation, and Verification (Ashbaugh 1999; Mnookin, 2008a). The examiner first inspects the two prints individually (analysis), then compares them relative to each other, looking for both similarities and differences (comparison). They then evaluate those similarities and differences to arrive at a decision about whether the prints match or not. In the final step, a second examiner independently validates the comparison. Mnookin points out that there is no formalized process for any of these steps. There is no method or metric for specification of which features should be used for comparison, nor any general measure for what counts as sufficient information to make a decision. Examiners rely on their experience and training rather than formal methods or quantified rubrics for making a decision. Despite the lack of a formal, standardized procedure, attempts have been made to formally describe and classify the kinds of features that might be found in a fingerprint.

Three types of features are commonly used to describe the information used for fingerprint comparison (for a complete discussion, see Maltoni, Maio, Jain, & Prabhakar, 2009). Level I features are global descriptors of ridge flow easily seen with the naked eye. The pattern in the central region (the "core") of the fingerprint can be classified as one of several common types: left- and rightward loops, whorls, tented-arches and arches. Deltas are triangular patterns that often occur on the sides of loops and whorls. A leftward loop and a delta are indicated by the yellow and green boxes respectively in Figure 1. Level I features are too common to be sufficient for identification, but they can be used for exclusion purposes as well as to guide inspection of the more detailed Level II and Level III features.

Level II features include *minutiae* such as ridge bifurcations and ridge endings. Level II features are found where fingerprint ridges and valleys split or end. *Minutiae* are highlighted in red circles in Figure 1. The uniqueness of fingerprints for identification purposes is largely due to the high variability in the existence and the relative positions of these features across fingers and individuals. Scarring, which occurs naturally with age and wear, can also add unique ridge patterns to a fingerprint. However, while scars can be used to compare the fingerprint found at a crime scene to that of a suspect in custody, they may not always exist in fingerprints on file that may predate the markings.

Level III features are the smallest fingerprint features used by some examiners for comparison. These include the positions of sweat pores and ridge thickness. Pores are indicated in light blue circles in Figure 1. The visibility of Level III features depends on the quality of the prints and examiners do not uniformly make use of them for comparison purposes.

Training may lead to an increase in the number of detailed local characteristics (*minutiae*) noticed by participants in a given print (Schiffer & Champod, 2007). With brief presentation times (under a second), when a viewer may not have enough time to compare many local features across two images in a matching task, experts utilized configural fingerprint information more efficiently than novices, focused on different information, and/or more effectively filtered out irrelevant information (Busey & Vanderkolk, 2005). What that information was, however, was not a primary focus of the research. Marcon (2009) had naïve observers rate "high quality" known prints and "low quality" latents for distinctiveness. Performance for categorizing pairs of prints as coming from the same source or a different source was higher for high-quality and high-distinctiveness images. These results suggest that performance suffers when fingerprint image quality is low, but do little to determine what makes a print low quality in the first place.

Figure 1. Depiction of various image features commonly identified by expert examiners. Red circles indicate *minutiae* (ridge bifurcations or endings); blue circles indicate pores (they appear as small white dots along a ridge); the yellow square indicates the delta; the green rectangle indicates the core, in this case a leftward loop.

Whereas the kinds of visual structures that may match or differ across fingerprints (core patterns such as whorls or loops and *minutiae*) have received some consideration, almost no analysis has been devoted to characteristics of image quality that may affect the fingerprint comparison task. These considerations apply mainly, but not exclusively, to latent prints. Intuitively, we would expect that a partial latent showing a small percentage of the full print, made on a surface unfavorable for extracting prints, and moved or smudged when the impression was made would present a more challenging matching problem than a clearer latent of larger area. For known prints, there is also variability in contrast, smudging, collection of excess media, and so forth that can affect the visual information available. There may also be relational variables involving the print pair: for example, comparing two prints of similar contrast may be easier than comparing a high-contrast known print and a low-contrast latent. Image processing measures extracted from latents and known prints, and the relations among them, may be useful for predicting the difficulty of a given comparison.

Predictor Variables

What properties of the images in fingerprint pairs are most important and informative in comparing fingerprints, and therefore most strongly predict matching performance? Although we relied on regression methods to provide answers to this question, it was important to develop, as inputs to the regression analyses, a wide variety of possible image characteristics that could be relevant. To generate such factors, we were guided by visual science, intuition, insights from fingerprint examiners, and prior work on image processing of fingerprints (e.g., Maltoni, et al., 2009), as well as the standard taxonomy of levels of pattern information in fingerprints. Some variables intuitively relate to the quantity of available information; for example, having greater print area available for

comparison might make comparisons more accurate. However, this view might well be oversimplified; quality of information might matter as much or more than total print size. We created and adapted several image processing techniques sensitive to smudging, missing regions, poor contrast, etc. In short, these algorithms were used to create variables with values for each print pair that likely relate to the visual information relevant to examiner performance.

As mentioned above, we hypothesized that difficulty would be a function both of the characteristics of the individual prints (the latent and the potential match) and also of the characteristics of the *pair.* Because known prints are obtained under relatively standardized conditions, they are subject to significantly less variability than latent prints obtained from crime scenes. Accordingly, we expected that more of the variability in visual information quality affecting fingerprint comparisons would be determined by characteristics of latent prints. An especially poor quality latent might be more difficult to assess than a higher quality one, all else being equal. However, we also believed that comparison difficulty would be a function of interaction effects between the latent and the known, not simply a function of the information quality and quality of each alone. We therefore developed quantitative measures involving both individual prints and print *pairs.*[2]

A general description and motivation for the image features we selected or developed is provided below. Except where noted, we assessed each predictor variable for the latent print and the known print. For many variables, we also derived a variable that expressed an interaction or relationship of the values for the latent and known print combined (such as the ratio of latent print area to the known print area, or the Euclidean sum of contrast variability for the latent and known print combined). Details about the procedures used to derive the measures are described below.

Total Area. This variable was defined as number of pixels in the fingerprint after the fingerprint was segmented from the background. Although machine vision algorithms exist that could have been used for determining the region of usable print image, those algorithms we examined were not as good as human segmentation, and different human observers in pilot work produced strong agreement. Accordingly, we segmented fingerprints from their surrounds by having human observers designate their boundaries. In general, we expected that larger areas, especially of latent prints, would provide more information for making comparisons.

While there are a variety of automated computer algorithms to segment a fingerprint from its background (Shen & Eshera, 2003), we opted to manually segment the images because, although the automated methods we tested worked well for most known prints and high quality latent prints, they failed for many low-quality latent prints. Since many of our latent prints were intentionally low quality (e.g., low contrast, smeared, etc.), the automated approach was not adequate. Furthermore, fingerprint technicians often manually specify the region in which a fingerprint is to be found, and so manually specifying the print region was not a great departure from standard procedures (observations from Los Angeles Forensics Lab). To calculate Total Fingerprint Area, a graphic user interface (GUI) was developed in MATLAB that displayed each image, one at a time. Two of the authors segmented all images by clicking and selecting points on the boundary of the print. A

[2] One of the anonymous reviewers of the draft report made the interesting observation that to look at the characteristics of 'pairs' rather than individual prints could be seen to violate the principle of separating the analysis phase from the comparison phase, of ACE-V a separation which many fingerprint analysts adhere to, and which has been recommended by some as a method for controlling the risk of bias (Expert Working Group on Human Factors in Latent Print Analysis, 2012). It is true that assessing the comparison exemplars in conjunction does not adhere to the principle of a complete separation of these phases. However, the purpose of this separation is as a mechanism to control cognitive bias. If a metric makes use of automated, objective measures for each print, that obviates the need for separation. To whatever extent a metric incorporates subjective dimensions of measurement, the reviewer's point would indeed have purchase.

polygon was fit through the points and the number of pixels within its boundaries was used as a measure of print area. Since each print was scanned at the same resolution, number of pixels is proportional to physical area.

Area Ratio. To relate the relative area of a latent to a potentially matching known print, we divided the area of the latent fingerprint by the area of the known print. Typically the known print, obtained under controlled conditions, presents a more complete image. Thus, *Area Ratio* relates to the proportion of known print information potentially available in the latent print. However, for non-matching prints, the area of the latent may be larger than that of the known print because of differing finger sizes. Occasionally, even for a matching latent and known print, the latent could be larger than the known print due to smearing. The ratio was therefore not strictly in the range [0,1] and cannot be considered a true proportion.

Image Intensity. We measured the mean and standard deviation of pixel intensity taking into account all of the pixels in each fingerprint image (with intensities scaled in the range of [0,255]). The mean intensity and standard deviation of intensity provide two related but different measures, sensitive to different image characteristics. Very dark images (low mean intensity) might indicate the presence of large smudges that produce large, dark areas. Low standard deviation in intensity would make ridges (transitions from light to dark) difficult to detect.

Block Intensity. The image was divided into 50x50 pixel regions and the average pixel intensity was computed within each region. The mean of the block intensities is the same as the overall mean *Image Intensity.* The standard deviation of these regional averages (*standard deviation of block intensity*), however, can provide additional information about variability in image intensity across the image. Low variability is indicative of many similar areas across the image, but does not provide information about whether those regions have low or high contrast (i.e., an all black image and an image with 50% white and 50% black pixels, evenly distributed across the image, would have low *Block Intensity* variability). When pixel intensities are not uniformly distributed across the image, variability of block intensity is high (i.e., some regions of the image are darker than others). For latent images, this may indicate the presence of a smudge or worse contact (lighter impression) in some regions of the image.

Deviation from Expected Average Intensity (DEAI). Intensity, as coded above, may be a useful predictor variable, but both intuition and pilot work led us to believe that it might not capture some significant aspects of intensity variations. We therefore developed a separate intensity measure – deviation from expected average intensity. In an ideal fingerprint image, one might expect approximately half of the pixels to be white (valleys) and half to be black (ridges). The expected mean intensity would therefore be half of the range, or 127.5.[3] The absolute deviation of the observed average from the expected average was computed using the following formula:

$$DEAI = -|mean\ pixel\ intensity - 127.5|$$

Using absolute value here ensures that deviations from the midpoint of the intensity range in either direction are scored as equivalent; the negative sign ensures that the measure increases as the mean pixel intensity approaches 127.5 (large deviations produce a large negative value of the measure).

Contrast. Michelson contrast was computed for each the segmented fingerprint. Michelson contrast is defined as:

$$Contrast = \frac{Maximum\ Intensity - Minimum\ Intensity}{Maximum\ Intensity + Minimum\ Intensity}$$

[3] Ridges, on average, are thicker than valleys so the expected average would be slightly lower since there would be more black pixels than white.

This contrast measure produces a value between 0 (least contrast) and 1 (most) by dividing the difference of maximum and minimum intensity values by their sum. Michelson contrast is typically calculated from luminance values. In our images, we calculate Michelson contrast from pixel intensity values, which is appropriate given that fingerprint images may be displayed on a variety of monitors with different Gamma coefficients.

Block Contrast. The preceding measure obtained the Michelson contrast for an entire image. We also computed contrast for smaller image regions – block contrast – by segmenting the entire image into 50x50 pixel regions. *Block Contrast* is defined as the mean across the blocks. To illustrate the difference between overall contrast and block contrast, the Michelson contrast of an entire image containing all gray pixels, except for one white and one black pixel, would be 1. *Block Contrast*, however, would be very low, since most regions of the image would have 0 contrast. If black and white pixels were distributed more evenly across the image such that they appeared in each block, then *Block Contrast* would be high. High values of the measure may indicate the presence of clear ridges and valleys in many areas of the fingerprint. A separate but related predictor was the *standard deviation of block contrast* across blocks. Small standard deviation values could indicate high information content throughout the image (*Block Contrast* close to 1 everywhere) or that the image was uniformly smudged (*Block Contrast* close to 0 everywhere).

Ridge Reliability. Orientation-sensitive filters were used to detect edges in the fingerprint image. The relative responses of these filters were then used to identify "high reliability" regions where ridge orientation was uniquely specified. The proportion of high reliability regions was computed, resulting in an overall reliability score for each print. Ridge Reliability ranged between 0 and 1, with larger values indicating a greater proportion of print area with well-defined ridge orientation. An additional, relational predictor was computed by taking the Euclidean sum of the *Ridge Reliability* for the latent and known print (*Ridge Reliability Sum*). Large values of this measure indicate a high proportion of regions with well-defined ridge orientation in both the latent and known prints.

Fingerprint images were histogram equalized in blocks of 75x75 pixels to 256 gray levels. Local ridge orientation reliability was then computed for each pixel in each latent and tenprint using the MATLAB function ridgeorient.m (Kovesi, 2000). ridgeorient.m first computes the pixel intensity gradient within a 10x10 pixel region centered on each pixel. For that region, the direction of maximum change in intensity was identified. The area moment of inertia was then computed about this direction. This is the minimum moment of inertia, while the perpendicular direction is the maximum. The ratio of minimum to maximum inertia was computed and subtracted from one. If the two moments are close to each other, then the gradient in the maximum and perpendicular directions is similar, meaning that there is little variation in intensity in any direction that region of the image and it is unlikely to contain an edge. This would yield a ratio close to one, and, when subtracted from one, a reliability value close to zero. In contrast, a clear edge would produce a large difference between the minimum and maximum moments of inertia and therefore a small minimum to maximum ratio. When subtracted from one, it would yield a reliability score close to one. This code is available for download (see Kovesi, 2000). The local reliability values at each pixel were then averaged across 50x50 pixel regions. Regions in which the average reliability exceeded a threshold of 0.45 were classified as reliable. The proportion of reliable regions in the segmented fingerprint image was the *Ridge Reliability* measure. This measure is bounded between 0 and 1 and corresponds to the proportion of the area of each print that contains reliable ridge orientation information.

Visibility of Cores and Deltas. Earlier we described global configurations – *Cores* and *Deltas* – that provide Level I information to fingerprint examiners. The fact that ridge flow in fingerprints tends to follow a circular pattern dictates that there will be some global core (a whorl, loop, or arch) at or near the center of each print. Likewise the transitions from global cores, especially loops and whorls,

to the circular ridge flow tends to give rise to deltas, triangular configurations (see Figure 1). As there will be only one core and at most a small number of deltas in any print, these serve as important reference points in making comparisons (Maltoni, et al., 2009). Unlike all of the other variables we used, which could take on a continuous range of values, *Cores* and *Deltas* are binary (either present or not).

A MATLAB-based GUI was developed and used by one of the authors to count the number of deltas present and whether or not the core was visible. Each print was also classified as left loop, right loop, whorl plain, whorl twin, arch or tented arch (or "unclear" if insufficient information was available for making a definitive classification). This interface is shown in Figure 2.

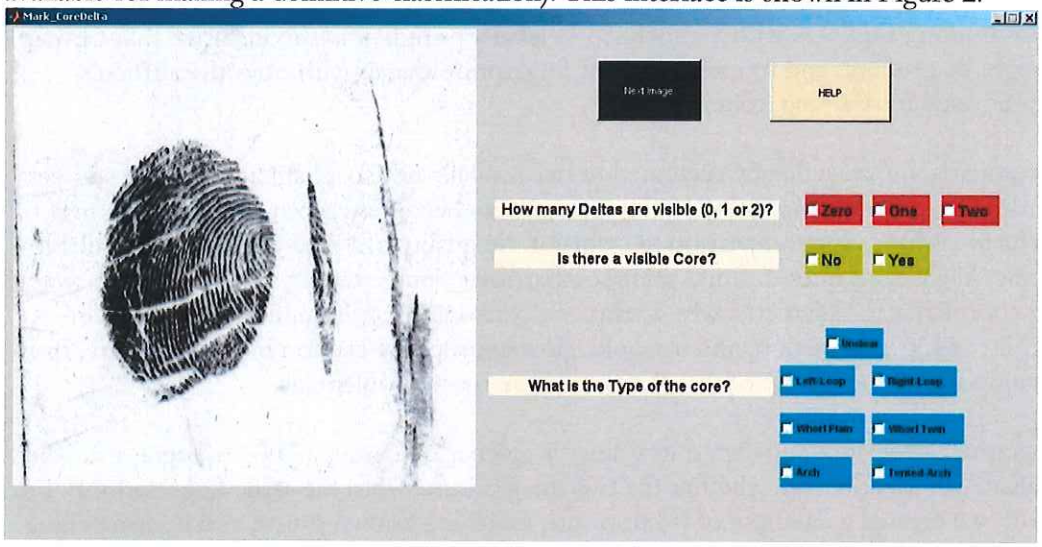


Figure 2. The interface used for counting deltas, marking the presence or absence of a core, and labeling the core type.

Relations Among Basic Predictors

To remove effects on regression coefficients of differing scales of various predictors, we standardized all continuous metrics by subtracting the mean and dividing by the standard deviation. Standardization made some measures that were strictly non-negative (like *Standard Deviation of Intensity*) take on negative values. As is often recommended in using regression methods (e.g., Neter, Kutner, Wasserman, & Nachtsheim, 1996), we also examined the features for collinearity and found that several predictors were highly correlated. For example, the mean and standard deviation *Intensity* measures were correlated (Pearson's r = -0.77 for latents and -0.44 for known prints). High correlation among predictors is an undesirable feature for regression models (Neter et al. 1996) because it makes it harder to assess the individual effect of those predictors. If two predictors had a correlation of greater than 0.5, we removed one of them. After removal, the variance inflation factor, a measure of collinearity, for all continuous metrics was less than 5, indicating that collinearity was sufficiently reduced (Chatterjee & Price, 1991; Booth, Niccolucci, & Schuster, 1994; Neter, et al. 1996).

In addition, we included two-way interactions between all predictors that applied to both a latent and known print. For example, in addition to the *Standard Deviation of Block Contrast* for the latent and known print, we included the interaction between the two terms. In addition to *Area Ratio* and *Ridge Reliability Sum*, these are *relational* predictors that encode something about the relative quality of information in a latent and known print.

Present Studies

The main goal of the project was to identify and quantify fingerprint image features that are predictive of identification difficulty and accuracy. Low quality prints recovered from crime scenes are often distorted, smudged, or contain only partial impressions. Experts may disagree whether the prints contain enough information or are of sufficiently good quality to make a determination (identification or exclusion / match or non-match judgment) or whether there is not enough information (i.e., the print is of insufficient quality) to make a judgment. It would be useful to have a quantitative way of assessing fingerprint image quality and comparison difficulty. Such a metric can be used to alert examiners when additional care is warranted (i.e., for a particularly difficult comparison), to caution examiners who are inclined to label a print pair as inconclusive that further examination might be prudent, and to create a set of fingerprint images with objective difficulty ratings that can be used for training examiners.

If fingerprint comparisons are generally accurate but occasionally not so, characterizing the sources of difficulty and the quality of information in fingerprint pairs becomes crucial. Ultimately, it may be possible to evaluate a fingerprint comparison in terms of the quality of visual information available in order to predict likely error rates in fingerprint comparisons. Such a metric would have great value in both adding confidence to judgments when print comparisons are uncomplicated in terms of having high quality visual information, and it would allow appropriate caution in cases that are, from an objective standpoint of the quality of visual information, more problematic.

In a typical fingerprint evaluation, an expert examiner is given a latent and a known fingerprint pair, which they evaluate for identity (i.e., whether the two images came from the same finger or not). For the present study, we created a database of latent prints, matching known prints, and non-matching prints retrieved from an AFIS database. Comparisons involving prints retrieved by AFIS resemble those performed in realistic settings where candidate matches are also generated by AFIS. Since the system attempts to find similar prints, the comparisons in our study may reveal error rates that are higher than that would occur if the non-matches were randomly selected, but would be more representative of real-world comparisons.

In all studies reported here, participants performed a two-alternative forced choice task in which they evaluated whether two fingerprint images came from the same source (matched) or from different sources (did not match). The latent print was not a cropped version of the known print; rather, the two prints were retrieved in different instances and the task was to determine whether the same finger created both. Images were presented side-by-side on the computer screen.

In addition to creating a fingerprint image quality metric, the project had several other objectives: (1) To create a realistic fingerprint image database with known ground truth about each pair (i.e., true matching prints and relatively close non-matches). (2) To add to the rather modest, albeit growing scientific literature investigating fingerprint expertise; that is, to assess objectively expert performance in terms of both accuracy in fingerprint identification and the image characteristics that related to performance. (3) To compare expert performance with novice performance, and in that manner quantify the degree of expertise. An additional benefit of studying novices was our ability to study how performance changes when a group of novices was made familiar with some characteristic visual features of fingerprints through brief training and to investigate how and to what extent this training changes cognitive strategies by altering the relative importance assigned to various visual characteristics in fingerprint pairs. Although not an original goal of the project proposal, this was a natural extension and resulted in an interesting finding.

We created a database of matching and non-matching fingerprint pairs that were used in all studies described in this report. The details of the database are described in the following section. Fingerprints were chosen to be a realistic example of the kinds of prints that are normally found in evaluation settings. Care was taken to attempt to generate pairs that varied in difficulty. First, this was important in order to ensure that sufficiently difficult comparisons were included to try to simulate difficult comparisons in the real world and potentially generate errors. Second, a range of difficulty allows for the database to be used in other settings, for example, as a training set from which examiners can select, easy, medium, and difficult comparisons. We used this database for several experiments reported below.

In Experiment 1, fingerprint examiners recruited from a forensics conference made match/non-match comparisons for a subset of print pairs from the database. There were several important differences between the experiment and typical comparison settings. Normally, examiners have the choice to label a print pair as "inconclusive", which means that the examiner deems that there is not sufficient information available to unambiguously say whether two prints come from the same finger or not. This creates the possibility of a different kind of error from saying that two non-matching prints are from the same person (false alarm) or saying that two prints from the same person are from different people (miss): incorrectly deeming that there is not enough information to make a conclusive evaluation when there is in fact sufficient information. In all experiments, we asked participants to provide difficulty and confidence ratings for each comparison. While this procedure is different from the operation of fingerprint analysis in normal forensic settings, it has two important advantages. Firstly, errors in this forced-choice framework likely have a more direct relationship to fingerprint quality. Second, we were able to examine the relationship between fingerprint information quality and confidence. This experiment was an important first proof-of-concept to demonstrate that under at least restricted settings, errors could be made. If it had turned out that experts made no mistakes given the constraints of the task, then there would have been little hope of artificially creating other situations in which errors could occur. To foreshadow some of the findings, several features of the fingerprint images were found to correlate with performance.

In Experiment 2, we used an overlapping subset of the fingerprints to test performance of novices. This served as a baseline comparison for examiner expertise. We expected that novices with absolutely no training would perform very poorly at this task since expertise requires extensive practice, in a same way that an amateur would have difficulty in classifying birds or determining whether an x-ray image contained evidence of cancer. However, novice performance seems to vary greatly depending on the type of study and materials, and can be as high as 75% for matching prints (e.g. Vokey, Tangen, & Cole, 2009; Tangen et al., 2011). It was therefore important to get performance measures for this particular set of images. One group of novices provided this performance baseline. A second group was shown a brief video that highlighted the kinds of image features used by experts in fingerprint matching and explaining how one might go about comparing fingerprints. It would be unreasonable to expect that watching a short video would drastically improve performance (otherwise, experts would not need such extensive training); however, the training video might cause novices to begin to use and be affected by the same information that experts use. We hypothesized that we might find that the kinds of features that were important for accurate performance for experts might receive a greater weighting or become more important for novices who watched a short video. For example, if it was pointed out that *minutiae* or ridge flow could an important factor in determining fingerprint identity, then perhaps measures like *Ridge Reliability* would become more important (meaning that performance would be higher for prints with greater values of this predictor) for the task. By comparing what predictors correlate with accuracy between experts and novices, we can examine differences between the two groups and identify which features may be most important to focus on.

Experiment 3 was an extension and validation of Experiment 1 with a different set of experts and an expanded set of tools in a substantially more realistic setting. Participants had access to a wide range of image processing tools to manipulate the images in the study, via an interface we built. They had unlimited time to make their comparisons. They also had the option of indicating that a particular comparison should have been deemed inconclusive. However, they still had to provide difficulty and confidence ratings, as well provide a best guess as to whether the prints were a match or non-match. The prints used in this study were selected based on predictions of difficulty from Experiment 1. Some of the prints in Experiment 3 were used in Experiment 1 and some were new. We expected to find generally comparable performance in this new group of experts to those tested in Experiment 1, but we did not know the extent to which the other manipulations (additional time and tools) would impact performance. If there was no difference in performance, then, moving forward, that would suggest that findings from experiments using simplified testing materials might be able to be extrapolated to more realistic settings; if there were very substantial performance differences, that would show that some of the simplifications of the sorts we took dramatically altered performance. We found a high correlation in accuracy for print pairs that were used in both Experiment 1 and Experiment 3, although there were some differences. The model was successful in predicting accuracy for many print pairs in Experiment 3, despite the differences between the two studies. There were several inconsistencies in predictions, however, particularly for print pairs that were labeled as inconclusive in Experiment 3. We explore some of the implications of these results and suggest further analyses and studies.

II. Methods

Database Creation

Fingerprints were collected from 103 individuals. Each individual first used a single finger to produce a clear, known print using ink as is done in police stations. Then, using the same finger, they touched a number of surfaces in a variety of ways (with varying pressure, smudges, etc.), to create a range of latent fingerprint marks that reflect those found in a crime scene. Professional fingerprint examiners who participated in the study reported that these prints were similar to those that they encounter in their everyday casework. The latent fingerprints were lifted using powder and were scanned at 500 dpi using the FISH system. Image dimensions ranged from 826 pixels in height to 1845 pixels and from 745 pixels in width to 1825 pixels. The latent prints that were created varied in clarity, contrast, and size. For each individual who contributed to the database, we collected a total of six prints – one known print and five matching latent prints. Across individuals we varied the fingers used. Each scanned fingerprint was oriented vertically and approximately centered.

To create the non-matching pair of prints, we did not want to randomly choose a known and a latent, as such pairs may be too obviously different. This would make the "non-match" decisions nearly uniformly easy, and would also, by default, indicate which were the "matching" pairs. Therefore, we obtained similar, but non-matching known prints by submitting the latent prints to an AFIS search. An expert selected from the AFIS list what he deemed as the most similar print. That enabled us to produce non-matching pairs that were relatively similar. The final database consisted of 1,133 fingerprint images – five latent prints from 103 fingers (515), 103 known prints that matched (103), and another 515 known prints for the non-match for each of the latents. Since we used an AFIS with a database from another country, it was practically impossible that a match would be in the database. Furthermore, the expert who selected the most similar print from the AFIS candidate list verified that each was a similar print, but not an actual match.

Experiment 1 – Experts at Conference

Subjects

Fifty-six fingerprint examiners (18 male, 35 female, three not reported) participated in the study. Forty participants self-reported as latent print examiners, three as known print examiners, ten as both, and three did not report. Years of experience were reported between the range of 1 and 25 years (Latent: Mean = 9.54, SD = 6.97 ; Ten-Print: Mean = 10.45, SD = 8.07). Twenty-seven participants reported being IAI (International Association for Identification) certified. 32 reported that their labs were accredited.

Participants were either directly recruited at the 2011 IAI Educational Conference or via a flyer sent out in advance of the conference. As incentive, participants were told they would be entered into a raffle to win an iPad 2. All participants signed informed consent forms prior to participating. As indicated above, some limited demographic information was collected, but it was stored separately from individual participant IDs such that the two could not be linked.

Apparatus

All stimuli were displayed on laptop computers with 17-inch monitors at a resolution of 1024 x 768 pixels. Stimuli were presented using a program accessed online; data were stored on the website's server. A screenshot of the program is shown in Figure 3.

Stimuli

Of the 1,133 fingerprint images, 200 latent and known print pairs were selected and used for the study; half were a match and half were a close non-match. Individual print metrics were computed for each image or image pair (see below) and prints were selected to (approximately) uniformly sample each feature space. Known prints were sampled without replacement, but multiple latent prints from the same finger were occasionally selected since each latent could be paired with a different known print image (the match or a close non-match). Print pairs were then grouped into batches of 20, each containing ten matches and ten non-matches. Latent prints from the same finger did not appear within the same batch.

Design

A group of experts made match / non-match judgments and provided confidence and difficulty ratings on a subset of 200 print pairs selected from a database of over a thousand fingerprint images. Two fingerprint images that were either from the same finger (match) or from two different fingers (non-match) were presented side-by-side. Images were presented on computer screens and were always oriented upright. Examiners had a maximum of three minutes to evaluate each pair of images. Performance was recorded for each print-pair tested.

Procedure

Participants were tested in a large room, seated at desks with individual laptop computers. Before data collection began, each participant was asked to sign a consent form, and then given written instructions detailing how the stimuli would be presented and the judgments they would be required to make. Participants were told that they would be asked to compare latent-known print pairs and determine whether they were matches or non-matches (without the option to choose "inconclusive" as a response). Participants were also told that they would be asked for confidence and difficulty ratings for each of their judgments. The instructions emphasized that this procedure was not

intended to replicate real-world conditions and that participants should simply try to maximize accuracy. Participants were also instructed to refrain from using any fingerprint examiner tools not provided by the experimenter, such as a compass.

When the experimental program was initiated, participants were asked to report their age, gender, years of experience, specialization, IAI certification, lab accreditation, and lab affiliation. Reporting this information was optional.

Next, the experiment began. On each trial, two fingerprints were presented side-by-side. The latent print was always on the left. A button in the top-left corner of each image window allowed participants to zoom in on each image individually. Fingerprint size was constrained within the bounds of each window, so that each print was always viewed through an aperture of 460 pixels by 530 pixels. The initial presentation of the images had them scaled to fit entirely in this window. A single level of zoom allowed participants to magnify the image. Participants could also translate each image independently within its window (both when the image was zoomed or unzoomed) either by dragging it with mouse or by using arrow buttons in the top-left corner of each image window. No other image manipulation features were available.

Participants made a match/non-match judgment by clicking a button at the bottom of the screen. Specifically, participants were asked: "Do these prints come from the same source or a different source?" Participants then made difficulty and confidence ratings by clicking on a Likert scale. The participants were asked: "How difficult is the comparison?" and "How confident are you in your decision?" On the Likert scales, "1" corresponded to least difficult / least confident and "6" corresponded to most difficult / most confident. Once all responses were recorded, an additional button appeared allowing the participant to advance to the next trial. Figure 3 shows a sample screenshot of the experiment.

Participants had three minutes to complete each trial. A message was given after two and a half minutes warning them that the trial will end in 30 seconds. If the full three minutes elapsed without a decision, the trial was ended, and the participant moved on to the next trial. After presentation of a set of 20 print pairs, participants were given a short break and asked if they wanted to complete another set of 20 comparisons.

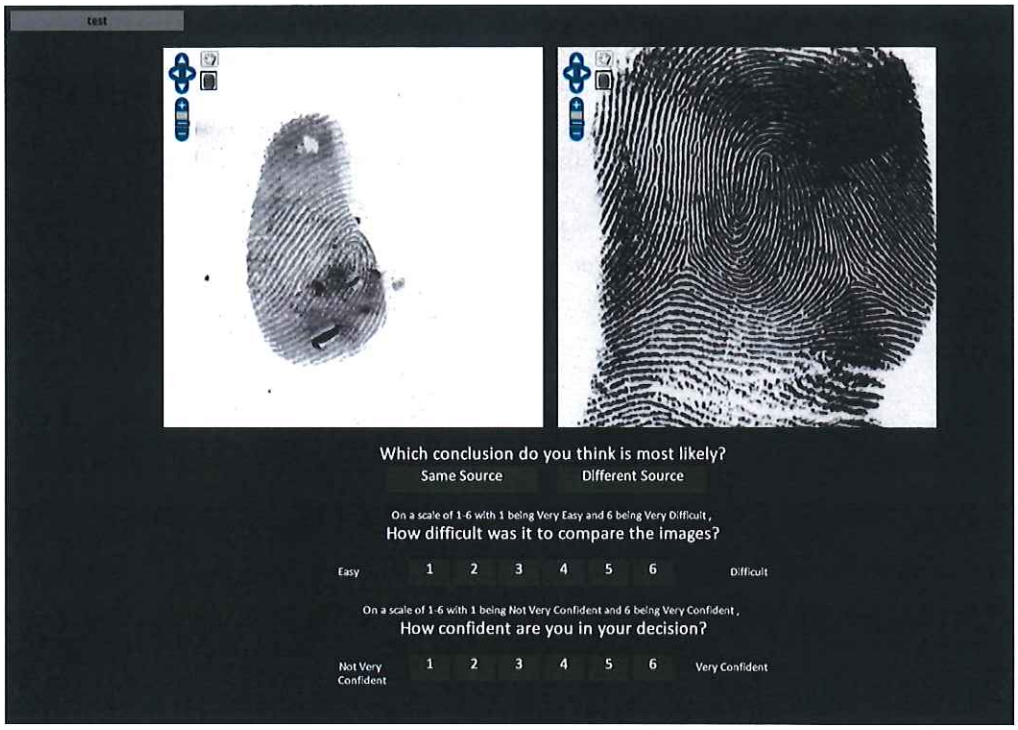


Figure 3. Screenshot of a sample trial from Experiment 1. Examiners could use the keys in the windows to change position or zoom level. Responses were made by clicking on the buttons shown in gray. Once all responses were provided, a button appeared allowing the user to advance to the next trial.

Each set of 20 print pairs contained ten match and ten non-match comparisons. The order in which print pairs were presented within a set was randomized across subjects. The sets were presented in a pseudo-random order so that approximately ten participants completed each set. Although the number of trials completed by individual participants varied based on their availability and willingness to do more comparisons, most participants completed two sets of prints (40 print pairs).

Experiment 2. Novices

Subjects

Participants were undergraduate students at University of California, Los Angeles, who participated in the experiment for partial course credit. 36 novices were randomly assigned to either the "no training" or the "training" groups, with 18 subjects per group.

Apparatus

All stimuli were presented using Matlab and routines from the Psychophysics Toolbox (Brainard, 1997). Displays were presented on one of three 16" x 12" ViewSonic Graphic Series G225f computer monitors in the UCLA Human Perception Laboratory, each with a resolution of 1024 x 768 pixels and a refresh rate of 60 Hz. The observer sat approximately 40 cm from the screen. Participants responded using a keyboard.

Stimuli

Fingerprint pairs (latents and known prints) were selected from the fingerprint database. As in Experiment 1, latents could be paired with a corresponding known print (match) or with a close non-match retrieved from AFIS (see Database Creation).

Subjects in the training group watched a 5 minute video ("How to Compare Fingerprints – The Basics") before beginning the experiment. The video described the fingerprint comparison process, identified cores and deltas and how they could be used in fingerprint matching, as well as other fingerprint features, such as *minutiae* and ridge counts. A sample comparison was performed in which *minutiae* were used to match two fingerprints.

Design

One hundred print pairs were selected randomly from the database. Each print pair came from a different individual. Half of the print pairs were matches and half were non-matches. Based on pilot data, novices went through comparisons fairly rapidly and could complete all 100 in approximately 40 minutes.

Prints were displayed side-by-side with the latent print always on the left-hand side of the screen and known prints on the right. Images were large, approximately 6 inches x 6 inches in size, although the size of the fingerprint within each image varied. Fingerprints were roughly 4 inches x 5 inches. The presentation order of comparisons was randomized across participants.

Procedure

Subjects sat at desks with computers in a well-lit room. On each trial, two fingerprints were presented side-by-side. The latent print was always on the left. Subjects responded whether the two prints were the same or different by pressing the Y or N keys on the keyboard. Each participant also made difficulty and confidence ratings. Participants were asked: "How difficult is the comparison?" (with 1 as easiest and 6 as most difficult) and "How confident are you in your decision?" (with 1 as least confident and 6 as most confident). Subjects responded by pressing a number key on the keyboard. Once responses to all three questions were entered, the participants could proceed to the next trial by a key-press. Participants were instructed that they should try to maximize accuracy. No other fingerprint examiner tools (e.g., a compass) were made available.

For the training group, subjects first watched an approximately 5-minute long YouTube video describing the fingerprint comparison process. Novices who received no training immediately started the experiment without watching any video.

The study began with a practice session of 6 comparisons on which subjects received feedback (correct or incorrect). After making a match response and submitting confidence and difficulty ratings, the two print images from the trial were shown again on the screen with the feedback printed above them to allow subjects to re-examine the images.

Experiment 3 – Experts with Advanced Tools

Subjects

Thirty-four examiners (16 male, 18 female; age range: 29-62), were recruited via personal contact. Twenty identified as latent examiners, one as a tenprint examiner, and 13 as both. Years of experience for latent examiners ranged from 1 to 36. Eight reported being IAI accredited. Thirty reported as coming from accredited labs or offices.

Apparatus

The experiment was conducted over a specially designed website that was a modification of the one used in Experiment 1. The basic structure was the same, including a login screen, consent form screen that included an electronic signature and a link to a downloadable pdf document that contained the consent information, a demographic form sheet that was optional, and the actual experiment page that displayed two fingerprint images. The welcome screen also included a password and login section. Passwords were e-mailed to users individually during recruitment and they were allowed to generate their own login names. Users were able to re-login as often as they liked and their progress was saved across sessions. The instruction screen was greatly expanded to include participation guidelines, system requirements, and image manipulation button control. All of these are described below.

Because subjects were allowed to complete the experiments remotely, no information about monitor size is available. In the instructions, users were asked to ensure that their monitors had a minimum resolution of 1200x720 pixels. Users were asked to click on a calibration button to adjust resolution and monitor brightness and contrast settings. Four shapes were shown and users were asked to adjust monitor resolution until all appeared to have equal side lengths with no distortion (pixel height and width should be equal). A brightness bar with 32 levels from black to white was shown below. Instructions stated that all 32 colors on the bar should be visible, with equal steps from bar to bar. In particular, users were instructed to adjust monitor contrast and brightness if the darkest bar was not seen or if there was a very large change in brightness between the final two bars. Users were expected to make these adjustments on their own. No feedback was provided and no measurements were taken by the website. Users were also instructed to use an up-to-date browser from among the following list: Firefox, Chrome, Safari, or Opera.

Unlike Experiment 1, the website had added functionality meant to reproduce some of the image processing features typically available to examiners in actual practice. Each fingerprint image (both the latent and known print) had a toolbar on the left-hand side with a 16 buttons. In addition, a navigation cross appeared within the boundary of the image that enabled users to pan across the image (up, down, left, or right) by clicking on the arrows of the cross. A zoom bar was located directly below it that allowed someone to step through four levels of zoom. All images began maximally zoomed out. A screenshot of this design is visible in Figure 4.

Figure 4. Screenshot of a sample trial from Experiment 3. The rest of the screen that included questions about difficulty and confidence ratings is not shown in this figure. Those questions would appear lower down on the page. The general layout was similar to that used in Experiment 1. In addition to the basic tools available in Experiment 1, an additional toolbar appeared to the left of each image to allow independent image manipulation. Hovering over an icon in the toolbar caused some hypertext to appear that described the tool. Clicking on the instructions button in the upper-right-hand corner provided access to detailed descriptions of each tool. A progress count indicating how many comparisons were completed appeared to the left of the instructions button.

A description of the image manipulations available in the toolbar appears below:

1) Zoom out completely (revert to original zoom level).
2) Free pan. By clicking on this button, a user would be able to manipulate the region of the fingerprint image that appears in the viewing window by clicking and dragging the image allowing different areas to become visible in the window. When maximally zoomed out, the entire image fit within the window.
3) Place new markers (on/off). Clicking on this button enabled marker placement. By default, the markers were green circles, however, the marker symbol and color could be changed. Markers remained in the correct positions on the image through zooming, panning, and rotation.
4) Change marker symbol. There were 5 marker symbols that could be used: circles, stars, crosses, triangles, and lightning bolts.
5) Change marker color. There were 4 marker colors: light green, dark green, red, and blue.
6) Erase marker. After clicking on this button, when the cursor hovered over a marker on the image, that marker was highlighted. Clicking on the marker removed it from the image.
7) Rotate image left (counter-clockwise) by 15 degrees.
8) Rotate image right (clockwise) by 15 degrees.
9) Reset rotation to original orientation.
10) Free rotate. Enable rotation one degree at a time by pressing the up and down keys.
11) Increase brightness of the image.
12) Decrease brightness of the image.
13) Increase contrast of the image.
14) Decrease contrast of the image.

15) Invert brightness of the image (black to white).
16) Undo all brightness and contrast manipulations (revert to values of original image).

An instructions button appeared in the top-right corner of the display. That could be used to re-access the screen shown at the start of the study. Image manipulations were not saved across login sessions. For example, if a user added markers to an image, but did not make a comparison judgment by clicking on the appropriate button and exited the experiment by closing the browser or logging out, the markers would not be visible when they logged back in.

Stimuli

Fingerprint pairs were selected from the database with the following constraints. Half of the examples came from Experiment 1. A subset was chosen that spanned a range of accuracies and included an equal number of matches and non-matches. This allowed us to validate the model on a new set of subjects. The regression model was then used to predict performance for the remaining print pairs that were not used in Experiment 1. Print pairs were selected to have a range of predicted accuracies from this set. In total, 120 pairs were chosen, 60 from Experiment 1 and 60 new pairs. Each set of 60 was composed of half matches and half non-matches.

Design

The design was similar to Experiment 1. Participants performed a two-alternative, forced choice (2AFC) task to determine whether two fingerprint images were a match (came from the same source) or not. There were 120 fingerprint pairs in the experimental set. The order in which they were presented was randomized across participants. Print images were presented side-by-side with the latent always on the left and the known print always on the right (see Figure 4). Each image had a manipulation toolbar to the left that allowed for a variety of image manipulations to be performed (see Apparatus section for details).

Procedure

Subjects were emailed a link to the experimental website as well as a password to access the site. On the welcome page, subjects read a brief description of the study and were asked to generate a username to use for accessing the site across sessions. Once a username and their password were entered, users were provided with a link to a pdf of the consent form and were instructed to download and read the form and to provide an electronic signature on the website. The following screen prompted users to provide some optional demographic information similar to Experiment 1. The next screen showed instructions for the experiment, the purpose of the study, and a description of the image manipulation tools as outlined above. The icon for each tool was shown, along with a description. Once the instructions were read, users continued to the actual experiment. Users made one comparison at a time and had to respond to all questions before continuing, similar to Experiment 1. Users were asked if the two images came from a same source or a different source and were asked to provide a difficulty and confidence rating for the comparison, similar to Experiment 1. In addition, and in distinction from Experiment 1, after making an identification or exclusion response, users were asked whether they would have marked the pair as inconclusive. If users logged out in the middle of a trial, the same trial would resume when they logged back in. However, all image manipulations that they had performed up to that point were reset. On a trial, users could adjust each of the two images independently with the manipulation tools. In addition, in this experiment, there were no time limits on each trial; examiners could take as much time as they wished. In all other respects – apart from the additional image manipulation tools, the absence of time limits, and the inquiry into whether the examiner would have selected 'inconclusive' if that had been an option – the procedure for this experiment was identical to that of Experiment 1.

Analysis Methods

Data Preprocessing

For the first experiment, if the examiner made a match/non-match judgment, but time expired before they could make difficulty or confidence ratings, the data were retained. If only difficulty and confidence ratings were provided, but a comparison judgment was not made before time expired, the trial was excluded from the analyses. Twenty such trials were excluded from the total of 2,312 comparisons (fewer than 1%) in Experiment 1. No other special preprocessing steps were undertaken for any other studies.

Descriptive Statistics and Correlations

For each experiment, average performance was computed for all comparisons, separately for match and non-match trials, separately for each individual, and for each print pair. Average difficulty, confidence, and response time ratings were also computed for each print pair and subject. Correlations were computed between accuracy, difficulty and confidence ratings, and response time. In Experiment 3, data were split by what tools were used and whether a print pair was rated as inconclusive or not.

Regression Analysis

We fit a crossed, logistic regression model in which print pair performance (1 = accurate; 0 = inaccurate) was crossed with expert and print identity. This is a type of mixed-effects model and is appropriate for analyzing these data for several reasons (Breslow & Clayton, 1993; Baayen, Davidson, & Bates, 2008). First, not every subject evaluated every print pair. A mixed-effects approach enables the examination of both the predictor variables and the "random effects" due to inter-subject differences (i.e., differences between expert performance and differences between evaluations of the same print pair by multiple experts). Second, a mixed-effects approach allows one to model individual item differences by fitting data from individual trials instead of aggregating across all presentations of an item (Dixon, 2008; Jaeger 2008). Differing levels of expertise and experience, as well as differences in comparison strategy and decision thresholds, could give rise to variability in participant performance independent of the fingerprint features. Variability across items could occur if some comparisons were easier than others irrespective of differences in measured image features. Including these sources of variability in the model allows us to test whether print comparisons and experts differed from one another, instead of assuming they were all equivalent and simply averaging across participants and items. Data were fit using the "arm" (Gelman & Su, 2013) and the "lme4" (Bates, Maechler, & Bolker, 2012) R packages for R version 2.15.2.

For each of i print-pair comparisons (items) and j experts (subjects), we define $y_{i,j}$ as

$$y_{i,j} = \begin{cases} 1 & \text{if print} - \text{pair } i \text{ is accurately classified (correct identification or rejection) by expert } j \\ 0 & \text{if print} - \text{pair } i \text{ is inaccurately classified (false alarm or miss) by expert } j \end{cases}$$

Accuracy for any particular print pair and expert, $\Pr(y_{i,j} = 1)$, was modeled with a logistic regression:

$$\Pr(y_{i,j} = 1) = \text{logit}^{-1}(X_{i,j}\beta + \text{printID}_i + \text{expertID}_j), \text{ for } i = 1, \ldots, 200, \; j = 1, \ldots, 56 \qquad (1)$$

where $X_{i,j}$ is a vector describing the features measured on a print pair, β is a vector of coefficients (the fixed effects; one coefficient for each feature), expertID_j is the expert-specific random effect, which allows the intercepts to vary across experts, and printID_i is the item-specific random effect. expertID and printID were normally distributed.

The regression equation can be rewritten and expanded as:

$$\text{logit}(\Pr(y_{ij} = 1)) = \beta_0 + x_{1\,ij}\beta_1 + x_{2\,ij}\beta_2 + \ldots + x_{n\,ij}\beta_n + \text{printID}_i + \text{expertID}_j \quad (2)$$

where n is the number of predictors. In this form, it can be seen that printID and expertID can be grouped with β_0 as intercept terms. Because printID and expertID are vectors, the equation reflects that each combination of print and expert has its own intercept term. It is this combined term (β_0 + printID$_i$ + expertID$_j$) that varies across experts and items. Multi-level modeling allows one to capture possible differences between individual subjects or test items without fitting a separate regression equation for each item (by applying a distribution over the terms that vary, in this case printID and expertID; see Gelman & Hill, 2007).

The parameter expertID is defined as:

$$\text{expertID}_j \approx \frac{\frac{n_j}{\sigma_\mu^2}}{\frac{n_j}{\sigma_\mu^2} + \frac{1}{\sigma^2_{\text{expertID}}}}\left(\bar{y}_j - \beta\bar{x}_j\right) + \frac{\frac{1}{\sigma^2_{\text{expertID}}}}{\frac{n_j}{\sigma_\mu^2} + \frac{1}{\sigma^2_{\text{expertID}}}}\mu_{expertID}$$

Where n_j is the number of print-pairs evaluated by expert j, σ_μ^2 is the within-expert accuracy variance, $\sigma^2_{\text{expertID}}$ is the variance among the average accuracies of different experts, $\bar{y}_j$ is the average accuracy for expert j, and $\mu_{printID}$ is the overall average accuracy across experts. From this equation it can be seen that expertID is a weighted average between the individual estimates of the intercept for each expert $\left(\bar{y}_j - \beta\bar{x}_j\right)$ and the average intercept across experts, $\mu_{expertID}$. When $\sigma^2_{\text{expertID}}$ is small, the right-most term dominates and the model approaches a regular regression model with a single intercept for all experts. When $\sigma^2_{\text{expertID}}$ is large, greater weight is placed on individual intercepts, and it is as if there is a separate regression model for each expert's data. expertID$_j$ is therefore a pooled estimate of the intercept term for each expert, taking into consideration across-expert differences in performance.

Each expertID is assigned the probability distribution

$$\text{expertID}_j \sim N(\mu_{\text{expertID}}, \sigma^2_{\text{expertID}}), \text{ for } j = 1, \ldots, 56$$

with the parameters of the distribution estimated from the data. One can see from this distribution that it has the effect of pulling the overall intercept closer to the average accuracy (μ_{expertID}) if there is little variability among experts (when $\sigma^2_{\text{expertID}}$ is small), and pushing toward individual regression equations for each expert when variance is large. The ratio of individual (within-examiner) and group (across examiners) variances is the intraclass correlation. It is defined as:

$$\frac{\sigma^2_{\text{expertID}}}{\sigma^2_{\text{expertID}} + \sigma_\mu^2}$$

When the intraclass correlation is close to 0 ($\sigma^2_{\text{expertID}}$ is small and σ^2_y is large), it indicates that differences between examiners contribute little to accuracy. Intraclass correlations close to 1 (large $\sigma^2_{\text{expertID}}$ relative to σ^2_μ) indicates that group differences contribute a lot to accuracy and that there is little variability within groups. Defining expertID$_j$ in this way allows the model to incorporate potential individual differences among experts. printID$_i$ is defined in a similar way.

Individual differences amongst experts may arise due to differences in experience, training, and other factors. These could manifest as different baselines of performance, or intercept terms in the model. All else being equal, one expert might do better with the exact same print pair than another expert. This variability is captured by the expertID term in the model. It is also possible to model

item-specific (in this case, print-pair-specific) effects; these are represented by printID. printID captures differences in print comparison difficulty inherent to individual print pairs and not related to the features used to predict print pair accuracy. In constructing a model, it is assumed that the error terms are uncorrelated; however, it is possible that print pair errors are correlated across participants. Inclusion of the item-specific term captures this potential non-independence (Baayen et al., 2008).

The regression model gave a predicted accuracy for each fingerprint pair. This was compared to the average observed accuracy. Model performance was measured as root mean squared error given by the following equation:

$$RMSE = \sqrt{\sum_{print\ pairs} (observed\ accuracy - predicted\ accuracy)^2}$$

RMSE values close to 0 indicate close agreement between observed and predicted accuracy across many print pairs; values closer to 1 indicate a poor fit. Further, we plotted observed versus predicted accuracy, fit a straight line to the data points, and computed R^2, a measure of linear fit.

Model Validation

In addition to creating models of accuracy, we fit similar models to difficulty and confidence ratings and response time data. Overlap in selected predictors with appropriate signs provides additional evidence for their importance. If, for example, *Area Ratio* was a significant positive predictor of accuracy, but was irrelevant for predicting difficulty and confidence ratings, we may have reason to be suspicious of its import.

We withheld 20% of the collected data from model fitting to use as a testing set. Models were fit on the remaining 80% of the data (the training set) and were then used to generate predictors for the withheld 20%. Performance was measured for both the training and testing sets. Testing sets are important to use to ensure that models are not over-fit to the specific sample.

Signal Detection Theory Measurements

In addition to basic accuracy information, one can distinguish between sensitivity and bias in subject responses. This is the basis of signal detection theory (Green & Swets, 1966). Sensitivity describes a sensor's ability to detect a signal. Once a signal is detected, the observer needs to make a decision about how to classify the signal, e.g., whether two prints were from the same source or not. Because sensors are subject to both external and internal noise, the exact same stimulus may elicit different responses across presentations. Sensitivity, *d'* (pronounced "dee-prime"), was computed with the following formula:

$$d' = Z(hit\ rate) - Z(false\ alarm\ rate)$$

Where Z is the inverse of the cumulative Gaussian distribution, *hit rate* is the proportion of "match" responses to match trials out of the total number of match trials, and *false alarm rate* is the proportion of "match" responses to non-match trials out of the total number of non-match trials. Values close to 0 indicate poor discriminability (inability to tell apart matching print pairs from non-matching pairs); higher values indicate better discrimination performance. For details, see, e.g., Green and Swets (1966).

Because of the high accuracy among experts found in other studies (e.g. Tangen et al., 2011), we expected their sensitivity to be very high, even for Experiment 1, which had time and tool constraints. We did not have a firm expectation for novices since reports of novice performance

were quite varied in terms of their performance. For example, Tangen et al. (2011) found accuracy to be around 75% for matches and around 50% for similar non-matches.

Bias was computed by calculating the log β. The measure can be thought of as the bias to respond "yes" or "no" in a forced-choice signal detection task. For the current study, the two alternatives can be thought of as "match" or "non-match" responses. It is also the log likelihood-ratio for a statistical decision test (see e.g., Wickens, 2002). The bias is computed by the following formula:

$$\log \beta = \log \frac{\varphi(\lambda - d')}{\varphi(\lambda)}$$

where φ(x) is the Gaussian density function, d' is the sensitivity, and λ is the decision criterion boundary given by –Z(false alarm rate).

A score of 0 indicates no bias. That is, no preference for saying "match" vs. "non-match" irrespective of one's discriminative ability (i.e., expertise). Deviations away from 0 indicate a preference toward saying "match" or "non-match". Positive bias scores indicate a more conservative decision criterion, a propensity to say "non-match" more often. Negative bias scores indicate a more liberal decision criterion, a propensity to say "match" more often.

We expected that experts would show a slight conservative bias, favoring "non-match" responses because of the high cost of making a false identification. Novices who received no training might not have the same associations with the fingerprint matching task and might show no bias. Tangen et al.'s study, however, indicates that there may be a bias toward saying "match". This would explain the significantly greater accuracy for matches compared to non-matches. Novices who watched the brief training video were made aware of the importance and difficulty of matching fingerprints and so might show a bias similar to experts or a reduction of the bias toward saying "match" shown by novices who did not watch the video.

III. Results

Experiment 1

Descriptive Statistics

Responses were aggregated across participants and prints. Overall accuracy (percent of correctly classified latent-known print pairs, averaged across subjects) was 91% (range: 8.3 -100%, SD 17%). Average accuracy was 86% for "match" trials (14% false negatives) and was 96.8% for "non-match" trials (3.2% false positives). Of the 2,292 comparisons, there were 200 errors, resulting in an overall error rate of 9.6%. Accuracy for particular print pairs ranged from 86% to 95%. There was some variability in performance among experts (range: 79-100%, SD 5%).

Non-matching trials include prints that do not originate from the same source; participants responded to a total of 1144 of these trials. Participants correctly labeled 96.8% of the non-matching trials as "no match" (correct rejections), and incorrectly labeled 3.2% of the non-matching trials as "match" (false alarms). In absolute terms, participants correctly labeled 1107 of the 1144 non-matching trials as "no match" (correct rejections), and incorrectly labeled 37 out of the 1144 non-matching trials as "match" (false alarms). Nineteen examiners made at least one false alarm, and twenty-seven of the non-matching fingerprint stimulus pairs caused at least one false alarm.

At the level of the stimulus, nineteen fingerprint stimulus pairs accrued one false alarm each; six accrued two false alarms each; and two accrued three false alarms each. At the level of the

participant, twelve participants made one false alarm; three participants made two false alarms; three participants made three false alarms; and one participant made ten false alarms.

Across all participants, 118 of the 200 print pairs produced 100% accuracy. Mean difficulty and confidence ratings for these pairs were 2.62 and 5.23 respectively, compared to ratings of 4.06 and 4.15 for prints that were misclassified by at least one participant. Of the 118 pairs that produced no errors, 72 were non-matches and 46 were matches. The lowest accuracy, 8.3% (1/12), was for a "match" print-pair. Average accuracy for each print pair is shown in Figure 5 sorted by increasing accuracy.

There was a significant difference between average ratings of difficulty for hits ($M = 2.95$, $SD = 1.58$), misses ($M = 4.57$, $SD = 1.25$), correct rejections ($M = 3.17$, $SD = 1.60$), and false alarms ($M = 5.16$, $SD = 1.04$), $F(3, 2278) = 69.51$, $p < .001$. Post-hoc comparisons revealed that all pairwise differences are significant at the 0.05 level (Bonferroni adjusted $p < 0.001$) except for the comparison of misses and false alarms (Bonferroni adjusted $p = 0.22$).

In assessing average confidence when participants were correct versus when they were incorrect (i.e., collapsing hits and correct rejections and misses and false alarms), there was a significant difference between average ratings of confidence for correct ($M = 4.96$, $SD = 1.41$) versus incorrect responses ($M = 3.30$, $SD = 1.57$), $t(2280) = -15.80$, $p < .001$.

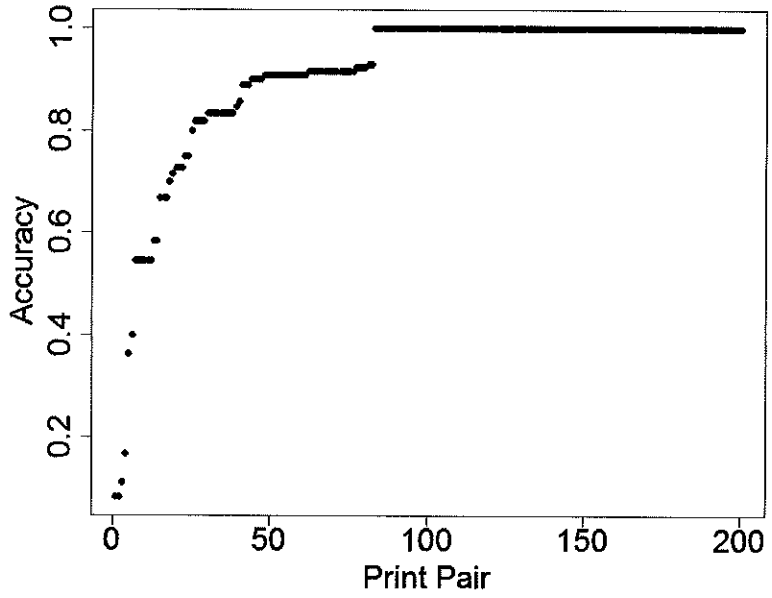


Figure 5. Sorted average accuracy for each print pair. Print pairs are numbered along the x-axis from 1-200 in order of increasing accuracy.

There was a significant difference between average ratings of confidence for hits ($M = 5.21$, $SD = 1.25$), misses ($M = 3.42$, $SD = 1.57$), correct rejections ($M = 4.74$, $SD = 1.50$), and false alarms ($M = 2.76$, $SD = 1.44$), $F(3, 2278) = 106.64$, $p < .001$. Post-hoc comparisons reveal that all pairwise differences are significant at the 0.05 level (Bonferroni adjusted $p < 0.001$) except for the comparison of misses and false alarms (Bonferroni adjusted $p = 0.06$).

There was a significant difference between average confidence ratings for trials in which participants responded "match" ($M = 5.12$, $SD = 1.34$) versus "no match" ($M = 4.57$, $SD = 1.57$), $t(2280) = -8.76$, $p < .001$. There was a significant difference between average confidence ratings for matching trials (M = 4.95, $SD = 1.44$) and non-matching trials ($M = 4.68$, $SD = 1.53$), t(2280) = -4.39, $p < 0.001$.

There was a significant difference between average difficulty ratings for trials in which participants responded "match" ($M = 3.03$, $SD = 1.62$) versus "no match" ($M = 3.35$, $SD = 1.63$), $t(2280) = 4.6867$, $p < .001$. There was no significant difference between average difficulty ratings for matching trials ($M = 3.18$, $SD = 1.64$) and non-matching trials ($M = 3.23$, $SD = 1.63$), $t(2280) = 0.8306$, *ns.*

For the seventy-four trials on which a particular examiner got a trial correct (hits + correct rejections) and for which at least two other examiners got incorrect (misses + false alarms), the average confidence rating was 3.51 ($SD = 1.75$) and the average difficulty rating was 4.59 ($SD = 1.32$). For the 837 trials on which a particular examiner got a trial correct (hits + correct rejections) and for which all other examiners got correct (hits + correct rejections), the average confidence rating was 5.00 ($SD = 1.34$) and the average difficulty rating was 2.86 ($SD = 1.53$). The difference between the confidence and difficulty ratings for the two sets were significant (confidence: $t(909) = 9.83$, $p < 0.001$; difficulty: $t(909) = -9.39$, $p < 0.001$).

There are many hits (438) and correct rejections (435) that the experts rated as not difficult (difficulty rating of 1 or 2) (total = 873 or 41.9% of the total number of correct responses). There were fewer hits (190) and correct rejections (257) that the expert rated as difficult (difficulty rating of 5 or 6) (total = 447 or 21.5% of the total number of correct responses). There were very few false alarms (1) and misses (9) that experts rated as not difficult (total = 10; 5.0% of the total number of incorrect responses). There were more false alarms (29) and misses (88) that the expert rated as difficult (total = 117; 58.8% of the total number of incorrect responses). This set of findings – showing that overall examiners have reasonably strong abilities to assess the difficulty of comparisons – offers interesting insights into examiners' metacognitive abilities, which we are currently in the process of analyzing further for an additional paper on the topic.

Correlations Among Dependent Measures

We measured the correlations of accuracy with the other three dependent measures. There was a strong negative correlation between average difficulty and confidence ratings ($r(198) = -0.91$, $p < 0.001$) and weaker correlations between average accuracy and confidence ($r(198) = 0.52$, $p < 0.001$), and between average accuracy and difficulty ($r(198) = -0.50$, $p < 0.001$). There was also a strong positive correlation between response time (RT) and difficulty ($r(198) = 0.71$, $p < 0.001$) and a negative correlation between response time and confidence ($r(198) = -0.59$, $p < 0.001$). Accuracy was highest and RT lowest for prints that were rated least difficult. Accuracy decreased and RT increased as print difficulty ratings increased. Excluding the 118 prints with 100% accuracy, the correlations between accuracy and confidence and difficulty were qualitatively weaker, but the difference did not reach significance. The full set of correlations is shown in Table 1.

Table 1. Correlations between dependent measures
All Fingerprint Pairs

	Accuracy	Confidence	Difficulty
Confidence	0.52***		
Difficulty	-0.50***	-0.91***	

Response Time	-0.48***	-0.59***	0.71***

Fingerprint Pairs with Accuracy < 100%

	Accuracy	Confidence	Difficulty
Confidence	0.36**		
Difficulty	-0.32**	-0.89***	
Response Time	-0.22*	-0.34**	0.45***

Note. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.5$

Regression Model

A cross, logistic regression model was initially fit to the entire dataset as described in the Analysis Methods section. A model was fit with all of the predictors (after removal of some to minimize collinearity). A likelihood ratio test showed that the model with the predictors fit the data better than a null model with only the random effects terms ($\chi^2(17) = 53.27, p < 0.001$).

Comparing a model that included the random expert effect (expertID) to one that did not, we found that the Akaike Information Criterion (AIC) was slightly smaller for the model that included the effect, but the Bayes Information Criterion (BIC) was smaller for a model that did not. Both of these measures are information-theoretic metrics of goodness-of-fit that take into consideration the overfitting the data with excess parameters. Qualitatively, a more parsimonious model that fit the data almost as well would have a smaller AIC and BIC (Akaike, 1973; Burnham & Anderson 2002). The fact that the criteria move in opposite directions when the model includes expertID suggests that any differences between the models should be treated with caution. A likelihood ratio test comparing the two models was significant ($\chi^2(1)=4.79, p<0.05$) (Zuur, Ieno, Walker, Saveliev, & Smith, 2009). expertID terms varied from between -0.52 ± 0.69 to 0.44 ± 0.77. All values of expertID were within two standard errors of zero. In terms of Equation 2, this means that β_0 + expertID was not reliably different from β_0. Based on these analyses, we felt justified in averaging across experts and ignoring between-expert differences in all subsequent modeling steps by removing the expertID term. This same analysis could not exclude the print-pair specific term, printID, which was retained in the model.

We simplified the model further by removing predictors (fixed effects) based on minimization of the AIC (Zuur et al., 2009). A likelihood ratio test revealed no statistically significant difference between a model that included all of the predictors and the reduced model ($\chi^2(11) = 9.55, p > 0.05$), indicating that the removal of predictors increased parsimony without significantly impacting predictive ability. Similar methods were applied to novice data from Experiment 2 and expert data from Experiment 3.

The model obtained for accuracy was:

$$Accuracy = logit^{-1}(3.385 + 0.798 * Delta\ (L) + 0.534 * Mean\ Block\ Contrast\ (K) - 0.471 * Area\ Ratio - 0.451 * SD\ Block\ Contrast\ (L \times K) + 0.419 * Sum\ of\ Ridge\ Measures + 0.334 * DEAI\ (L \times K) + printID)$$

Where L and K indicate whether the predictor applies to a latent or known print image respectively, and LxK indicates predictors that apply to print pairs. printID is the item-specific, random effect. The parameters of the fitted model are shown in Table 2. All predictors were significant (Wald's z, ps < 0.05), except for Delta (L) and DEAI (LxK) which were marginally significant ($p = 0.054$ and p

= 0.053 respectively)[4]. To get a more intuitive notion of model performance, we used the predicted proportions from the logistic regression as estimates of average performance across experts. The resulting fit was very good (R^2_{adj} = 0.91). We also computed the root mean squared error (RMSE) by taking the sum of the squared differences between predicted and observed values. Values closer to 0 indicated better performance. The error for the fitted model ($RMSE_{model}$ = 0.06) was lower than for a null model that only included the printID random effect ($RMSE_{null}$ = 0.18).

Table 2. Predictors for accuracy model.

Fixed Effects	Coefficient Estimates	Standard Error	z
Intercept	3.385	0.197	17.167***
Delta (L)	0.798	0.415	1.923
Mean Block Contrast (K)	0.534	0.164	3.268**
Area Ratio	-0.471	0.156	-3.010**
SD Block Contrast (LxK)	-0.451	0.128	-3.530***
Ridge Sum	0.419	0.154	2.715**
DEAI (LxK)	0.334	0.173	1.938

Random Effects	Variance
printID	2.154

Note: *** p < 0.001, ** p < 0.01, * p<0.05. p-values are reported here, but should be interpreted with caution. They were not used for model selection (see Footnote 2). Estimates are arranged by coefficient magnitude in descending order (see text). L – latent, K – known print, LxK – interaction.

Validation of the Regression Model for Accuracy

The dataset was then split into training and testing sets. First training on the full dataset was used as a check to make sure that the model could fit at all. If it had failed to fit on the full dataset, there was no point in training on a subset of the data. The training set contained 180 (90%) of the print pairs (2063 individual observations), and the testing set contained the remaining 20 print-pairs (10%, 229 observations). The testing set print pairs were a representative sample of the overall dataset, containing 12 pairs with perfect accuracy and 8 pairs with less-than-perfect accuracy. This was important in order to ensure that the training set did not have too few pairs with low accuracies (there were only 24 pairs in all with average accuracies below 80%). We replicated the model selection procedure for data only from the training set. The same predictors were selected with comparable coefficients, except for Delta (L) which was replaced with Core (L). For both the full and training datasets, the coefficients for these two predictors, Delta (L) and Core (L), were not significantly different from zero and were within two standard deviations of zero. Nevertheless, they could not be excluded based on the selection procedure described above. The fit of the model to the training set was comparable to that of one on the full set (R^2_{adj} = 0.89, $RMSE_{train}$ = 0.07). The results of fitting on the full set were therefore not likely due to overfitting.

We used this regression model fitted to the training set to predict accuracy for the withheld training set of 20 print pairs. Less variance could be accounted for the testing set than for the training set, suggesting some amount of overfitting (R^2_{adj} = 0.64). The error, however, was comparable between the training and testing sets ($RMSE_{test}$ = 0.07). The model's predictions are shown in Figure 6.

[4] *p*-values for the Wald statistic in unbalanced, mixed-effects data are difficult to define due to difficulty in determining the appropriate degrees of freedom and therefore should be interpreted with caution (Agresti, 1996, 2002; Baayen et al., 2008).

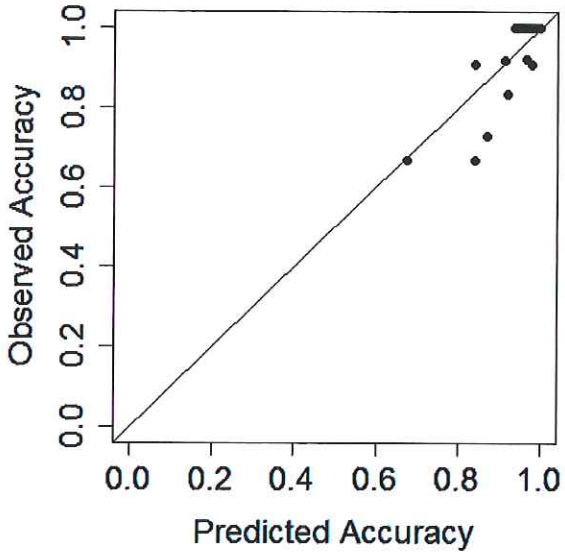


Figure 6. Model predictions of average accuracy for 20 test print pairs plotted against observed average accuracy.

As a secondary assessment of model performance, we used the model to predict whether at least one expert made an error on a print pair. We divided the set of print pairs into two classes: those that had 100% accuracy (perfect pairs) and those that had less than 100% accuracy (non-perfect pairs). A naïve classification strategy not based on the model and that assumes no errors are ever made would have a classification accuracy of 107/180 or 59%. Using the model fitted to the training set, we parametrically varied a classification threshold such that print pairs with a predicted accuracy greater than or equal to that threshold were classified as perfect pairs and those below that predicted accuracy were classified as non-perfect pairs. A threshold setting of 94% resulted in the best classification performance of 164/180 or 91% correctly labeled pairs.

The classification procedure described above was repeated for predictions generated for each left out (testing) pair using the threshold optimized on the training set. 75% (15/20) of the pairs were correctly classified as either having perfect (9/15) or non-perfect accuracies (6/15). The classifier was slightly better at correctly identifying print pairs that had at least one error than those that were perfect: 3 perfect prints were misclassified as having an error and 2 non-perfect pairs were misclassified as perfect.

Difficulty Ratings

Difficulty ratings showed a reliable negative correlation with accuracy (see Descriptive Statistics, above), indicating that experts had reasonable metacognitive awareness (i.e., print pairs that were thought to be difficult tended to have low accuracy across experts). Accuracy for trials with a difficulty rating greater than 3 (on a scale of 1 to 6) was 84% compared to 91% for all comparisons. We compared the fitted model from the previous section to one that also included difficulty rating as a predictor. The resulting model had significantly better goodness of fit than the model from the preceding section that did not include it as a predictor ($\chi^2(1)=81.1$, $p<0.001$, $RMSE_{model+difficulty} = 0.05$, $R^2_{adj} = 0.95$).

We added difficulty rating as a predictor for the regression model applied to the training set described above. Predictive performance on the testing set was worse (decreased R^2) than when the difficulty rating was not included. However, classifier performance on the testing set was slightly improved, with 85% (17/20) of the pairs classified correctly. One perfect print was misclassified as non-perfect, and two non-perfect prints were misclassified as perfect. The discrepancy between the relatively worse regression fit and the improvement in classifier performance is due to two non-perfect print pairs that had a predicted accuracy that was much lower than their true accuracy. These were classified correctly as non-perfect, but contributed significantly to the error.

The inclusion of difficulty ratings in applications of this model must be made with caution. All other measures capture objective features of the fingerprint image, while difficulty ratings are subjective and therefore may vary across individuals and rely on the good faith of the raters. Therefore, while difficulty rating may be informative to include, in subsequent models we opt to exclusively deal with objective factors. We return to this point in the discussion.

Regression Analysis of Other Dependent Measures

Difficulty ratings, confidence ratings, and response times were reliably correlated with accuracy and so ought to also depend on print pair information content. If similar features are predictors for many measures, then they are likely capturing something important about the fingerprint images. Here, we fit models of the other dependent measures to the training dataset as a further validation step: the importance of particular image features as valid predictors of accuracy is bolstered if those same features are shared in models of other dependent measures.

Unlike accuracy, response time varied greatly across experts, with some taking much longer times on comparisons that others evaluated fairly quickly. There are several possible reasons for this variability. Less experienced examiners may take longer to come to the same conclusion than a seasoned examiner (a perceptual fluency that comes with expertise; see Kellman & Garrigan, 2009). Some subjects may have completed the comparison quickly, but then taken time to deliberate confidence and difficulty ratings since response time was recorded only once all answers were given, and not when the subject selected "match" or "non-match". Also, the self-confidence of the examiners in their abilities may have affected response time. Only a small component of the variability in response time was likely to be due to differences in attention or interest since such differences would presumably have led to greater variability in accuracy, which was not observed.

We fit a linear, mixed-effects model to normalized response time data for the training set following the same model selection steps as for the accuracy model in the preceding sections. Due to the variability in response time across experts, the random effect of expertID was retained in the model. The results of the regression are shown in Table 3. Three features, Core (L), Mean Block Contrast (K), and SD Block Contrast (L) were found to be predictive of response time using the same model selection procedure that was used for the analysis of predictors of comparison accuracy. The latter two predictors were also selected in models of accuracy (SD Block Contrast as part of an interaction term). Visibility of cores instead of deltas was selected as a predictor of response time. Interestingly, it also appears as predictor when the model is fit to a testing set. Visibility of a core might make it simple to compare latents and known prints: if the cores do not match then no further comparison is required, so a comparison can be made quickly. Absence of a core could also make it difficult to orient the latent and known prints, since, as noted earlier, these features could act as landmarks for orienting two prints during comparison.

Linear mixed-effects models were also fit separately for difficulty and confidence ratings. Like response time, there was a great deal of inter-subject variability for both measures. Variability in

confidence and difficulty ratings may be due to differences in degree of expertise and self-confidence in the task. Variability in ratings may also be due to differences in interpretation of the rating task and therefore in response strategy. One expert, for example, responded with maximum confidence to all comparisons, saying to the experimenter that in real-world situations an expert would be 100% confident or rate a comparison as inconclusive.

Table 4 contains the coefficient estimates for the model of difficulty rating. As in the model of accuracy, Ridge Sum, Area Ratio, and Core (L) were selected as predictors. Similar to response time, difficulty was also negatively correlated with accuracy, so the regression coefficients have opposite signs to those in the accuracy model. In addition, visibility of Cores in the known print and the interaction of the Core terms were also selected. Delta (L) appears in this model as well as in the model of accuracy.

Table 4. Predictors for difficulty rating model.

Fixed Effects	**Coefficient Estimates**	**Standard Error**	**T**
Intercept	2.748	0.301	9.121***
Core (L x K)	-2.104	0.722	-2.913**
Core (L)	1.719	0.705	2.437**
Core (K)	0.935	0.324	2.883**
Delta (L)	-0.778	0.191	-4.082***
Ridge Sum	-0.207	0.079	-2.631**
Area Ratio	0.202	0.078	2.571**

Random Effects	Variance
printID	1.076
expertID	0.301

Note: *** $p < 0.001$, ** $p < 0.01$, * $p<0.05$. Estimates are arranged by coefficient magnitude in descending order (see text). L – latent, K – known print, LxK – interaction.

A similar model was fit for confidence ratings. The results are shown in Table 5. Identical predictors with comparable magnitudes were selected as for the difficulty rating model. The coefficients have opposite signs since high difficulty ratings correspond to low confidence ratings. Because difficulty and confidence are so strongly correlated (-0.91), it is not surprising that the exact same predictors are selected for in both models.

Table 5. Predictors for confidence rating model.

Fixed Effects	**Coefficient Estimates**	**Standard Error**	**t**
Intercept	5.248	0.247	21.255***
Core (L x K)	2.034	0.564	3.604***
Core (L)	-1.644	0.551	-2.983**
Core (K)	-0.920	0.253	-3.631***
Delta (L)	0.581	0.149	3.899***
Area Ratio	-0.162	0.062	-2.647**

Ridge Sum	0.155	0.062	2.517**

Random Effects	Variance
printID	0.616
expertID	0.488

Note: *** p < 0.001, ** p < 0.01, * p<0.05. Estimates are arranged by coefficient magnitude in descending order (see text). L – latent, K – known print, LxK – interaction.

Experiment 2

Descriptive Statistics

Overall accuracy for untrained novices was 53%. There was a significant difference for the average accuracy for "match" trials (62%) and "non-match" trials (45%, $t(49) = 3.51$ $p < 0.001$). Accuracy, averaged across prints for individual participants, ranged from 42% to 62% (M = 53%; SD = 5.8%). Of the 1800 comparisons, there were 838 errors, resulting in an overall error rate of 47%.

Overall accuracy for trained novices – and by 'trained' we mean only exposure to the short video presentation about fingerprint evidence and how it functions – was 54%. There was no significant difference ($p > 0.05$) for the average accuracy for "match" trials (54%) and "non-match" trials (54%). Accuracy, averaged across prints for individual participants, ranged from 43% to 75% (M = 54%; SD = 7.5%). Of the 1800 comparisons, there were 826 errors, resulting in an overall error rate of 46%.

The five highest accuracies for trained novices were 58%, 59%, 60%, 64% and 75%. The five highest performing untrained novices had accuracies of 57%, 60%, 61%, 61%, and 62%. Accuracy and rating scores are depicted in Figure 7. Expert scores from Experiment 1 are included for comparison.

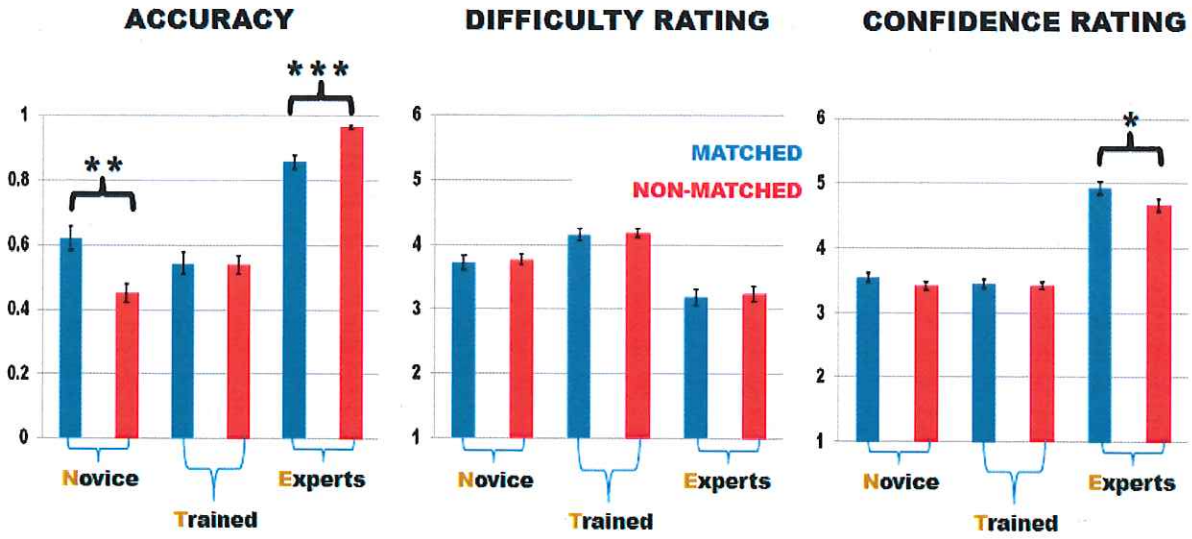


Figure 7. Mean accuracy, difficulty and confidence ratings for untrained novices, trained novices, and experts. Data are split by matching and non-matching comparisons. Error bars are standard errors. *s indicate significance levels of independent *t*-tests between matching and non-matching comparisons. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

For both groups of novices, there was a negative correlation between difficulty and confidence ratings ($r(198) = -0.87$ and $-0.85, p < 0.001$, with and without training respectively), and a weaker correlation between accuracy and confidence ($r(198) = 0.31$ and $0.33, p < 0.001$, with and without training) and between accuracy and difficulty ($r(198) = -0.41$ and $-0.31, p < 0.001$, with and without training respectively). It is interesting to note that the latter correlation is higher for novices with training than without training, perhaps indicating that they were beginning to appreciate the visual features that make matching print pairs harder.

The variance for experts' accuracy was less than that for either group of novices. This reflects the near-ceiling performance of the experts. The average difficulty rating was lower for experts (3.2/6) than either group of novices (untrained: 3.7/6, $t(30) = 3.50, p < 0.01$, trained: 4.2/6, $t(27) = 5.81, p < 0.001$). There was also a significant difference in difficulty ratings between the two groups of novices ($t(33) = 2.25, p < 0.05$). Confidence ratings of experts (4.8/6) were also higher than those of novices (untrained: 3.4/6, $t(37) = 7.26, p < 0.001$, trained: 3.6/6, $t(39) = 7.18, p < 0.001$). There was no significant difference in confidence ratings between the groups of novices. For the experts there was a significant difference between confidence ratings for match trials (4.9/6) and for non-match (4.7/6; $t(99) = 1.98, p < 0.05$)). No such asymmetry was found for the novices.

Signal Detection Measures

To assess participants' sensitivity in discriminating matches and non-matches, we submitted accuracy scores from the assessed print pairs to a signal detection analysis (Green & Swets, 1966). The average sensitivity for the expert group ($d' = 2.64$) was much higher than for the novices ($d' = 0.19$). There was no significant difference between the average sensitivity of untrained ($d' = 0.17$) and trained ($d' = 0.21$) novices. Despite low average sensitivities, the maximum sensitivity was 0.63 for untrained novices and 1.36 for trained novices. However, only 2/18 trained novices had sensitivities higher than the maximum untrained novice sensitivity.

Response bias ($\log \beta$) was computed for novices and trained novices. Mean bias (averaged across subjects) was 0.01 and 0.04 respectively. There was no significant difference between the two groups ($t(34) = -1.19, p = 0.24$). Average bias for the six highest performing untrained novices was slightly liberal (-0.06), while the average bias for the two highest performing trained novices whose sensitivity was greater than the maximum sensitivity of untrained novices was slightly conservative (0.12), but the difference between the two was not statistically significant ($t(6) = -2.13, p = 0.073$) perhaps because there were so few trained novices with high sensitivities.

Regression Analysis

The same crossed, logistic regression model was fit to the novice data as was used for experts in Experiment 1. Similar procedures were followed to remove variable and simplify the model. The results are shown in Table 6 with the coefficients from the fit to the expert data included for ease of comparison. Ridge Sum, Mean Block Contrast (K), SD Block Contrast (LxK), DAEI (LxK), and visibility of Cores (K) were selected as significant predictors for novices. Three predictors, Delta (K), Ridge Sum, and DEAI(K), were selected in the trained novice model.

Table 6.

	Expert	Untrained Novice	Trained Novice
Fixed Effects	Coefficient Estimate		

	(Standard Error)		
Intercept	3.385 (0.197) ***	0.521 (0.326)	0.628 (0.231) **
Area (K)			
Area Ratio	-0.471 (0.156) **		
Delta (L)	0.798 (0.415)		
Delta (K)			-0.523 (0.246) *
Ridge Reliability (K)			
Ridge Sum	0.419 (0.154) **	0.297 (0.112) **	0.312 (0.096) **
Mean Block Contrast (K)	0.534 (9.164) **	-0.540 (0.112) ***	
SD Block Contrast (L)			
SD Block Contrast (LxK)	-0.451 (0.128) ***	0.194 (0.085) *	
DAEI (K)			-0.253 (0.099) *
DEAI (LxK)	0.334 (0.173)	-0.213 (0.101) *	
Core (L)		0.463 (0.251)	
Core (K)		-0.752 (0.373) *	
Core (LxK)			
Random Effects			
	Variance		
Print Pair	2.154	0.809	0.658
Subject			0.077

* $p<0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 6. Coefficient estimates for the three groups of subjects: experts, untrained novices, and trained novices, and for a high-performing subset of the trained novices. L – latent K – known print L*K - interaction. Because a model selection procedure was used to select the most parsimonious model, some parameters do not appear in all models. Fixed effects appear at the top of the table and random effects appear at the bottom. For random effects, the estimated variance is specified. *p* values correspond to significance tests on Wald statistics for each predictor, which are not shown in this table. For mixed-effects models, it is difficult to determine the appropriate degrees of freedom, so *p* values should be interpreted with caution. Instead, it may be more informative to examine whether predictor coefficient estimates are within two standard errors of the 0. Predictors are sorted first in descending order of coefficient magnitude for experts and then by L, K, and L*K.

The root mean squared error (RMSE) was used as a measure of model performance on a withheld dataset of 20% of the prints similar to Experiment 1. RMSE was computed by making individual accuracy predictions for each print pair and then comparing this predicted average accuracy to the observed average accuracy. Point estimates of the predictor coefficients and random effect terms were used. RMSE for the expert, novice, and trained novice testing sets were 0.07, 0.25, and 0.21, respectively. The larger RMSEs for both groups of novices indicate poorer model fits. Regression predictors can still be interpreted as important contributors in predicting accuracy, but the model should be interpreted with caution. The poor fit is not surprising given near-chance performance for both groups of novices. However, it is interesting that despite these worse prediction results, a different, almost completely non-overlapping set of predictors is selected for in the trained novice

model, and that the prediction performance is slightly improved relative to the untrained novice model.

Experiment 3

Descriptive Statistics

Thirty-four examiners made a total of 1646 comparisons. Each print pair was evaluated by a minimum of seven distinct examiners. Average accuracy was 94.84%. Performance on matches was 90.00% while performance on non-matches was 99.75%. The lowest accuracy for any print pair was 10%. There were three print pairs out of 120 with an average accuracy less than 50%, three print pairs with an average accuracy between 50% and 75%, and 16 print pairs with an average accuracy between 75% and 100%. 98/120 print pairs had perfect accuracy.

Average examiner accuracy was 95.03%. The lowest accuracy was 81.82%, the highest was 100%. Four examiners computed fewer than 10 comparisons, but none made any mistakes. Eighteen examiners completed between 10 and 50 comparisons with an average accuracy of 93.55%. Twelve examiners completed more than 50 comparisons with an average accuracy of 95.6%.

Of the 1646 total comparisons, 126 were labeled as inconclusive, of which 73 were matches and 53 were non-matches. Average accuracy for prints labeled inconclusive was 76%; average accuracy for prints not labeled inconclusive was 96.38%. Average difficulty rating for inconclusive prints was 4.62; average difficulty rating for non-inconclusive prints was 2.27. For pairs that were labeled inconclusive by any examiner, an average of 23% of examiners labeled those prints inconclusive. At most, 7/9 examiners rated a particular pair inconclusive. Of the 42/120 pairs that had at least one examiner label inconclusive, five had 50% or more of examiners agree that they were inconclusive with an average accuracy of 59.78%. The remaining 37 comparisons had fewer than 50% of the examiners that rated them as inconclusive and had an average accuracy of 90.19%.

Half of the print pairs used in this experiment were also used in Experiment 1. Performance was strongly correlated across the two experiments on that subset of comparisons (Spearman's rho = 0.45, $p < 0.001$).. The accuracies for the two experiments are shown in Figure 8. Qualitatively, accuracy for many pairs was similar across both experiments. However, for several pairs there were marked differences. For two pairs, for example, accuracy in Experiment 1 was close to 50%, but was near 100% in Experiment 3. Another print pair had an accuracy of near 10% in Experiment 1 and an accuracy of approximately 55% in Experiment 3. We have not yet examined the kinds of tools that were used with each of these comparisons (see Conclusion for planned future analysis of these data).

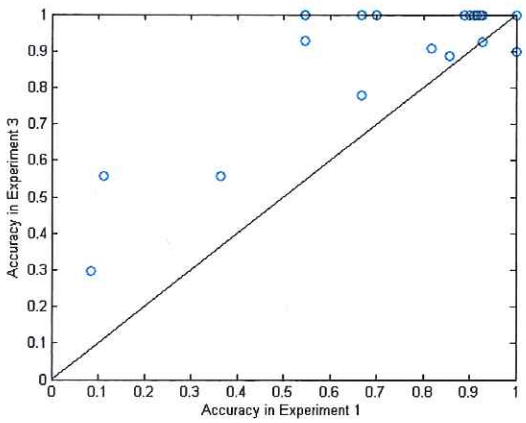


Figure 8. Average accuracy for the 60 print pairs in both Experiment 1 and Experiment 3.

Response Time

Minimum response time was 2.4 seconds. Maximum response time was 91.5 minutes. Average response time was 2.58 minutes. Average accuracy for comparisons that took less than a minute (21.43% of all comparisons) was 97.12%; average accuracy for comparisons that took more than a minute was 93.02%. Since users could leave the experiment window open (there was no time-out), the response time does not necessarily reflect the amount of time spent evaluating the fingerprints.

Tool Use

Across all comparisons, 82.38% made some use of the tools. Average accuracy for comparisons involving tool use was 95.86%, while average accuracy for comparisons without tool use was 94.62%. Average difficulty rating for comparisons on which tools were used was 2.61; average accuracy for comparisons without tool use was 1.72.

Minutiae were marked on 41.86% of the comparisons. On average, 3.74 *minutiae* were marked per comparison. Average number of marked *minutiae* was 4.42 for comparisons rated inconclusive and 3.68 for comparisons not rated inconclusive. Accuracy was 95.3% for comparisons with no *minutiae* marked and 94.19% for comparisons with at least one marked. Average difficulty rating for comparisons with no *minutiae* marked was 2.17 and 2.84 for those with at least one marked.

For the other tools, 79.65% of comparisons had the zoom tool used, 12.64% used rotation, 24.30% had a brightness or contrast adjustment. In all cases, average difficulty was rated as higher for comparisons that had tool use compared to those that did not (2.61 vs. 1.85, 2.90 vs. 2.39, 3.15 vs. 2.23 for each of the tools respectively).

Regression Analysis

The model fit in Experiment 1 was used to predict accuracy data from this experiment. The predictions were based on the un-edited images, i.e., it did not take into account if examiners used a tool to alter image properties like brightness or contrast. Since the model takes those features as inputs, the model predictions need to be interpreted with caution. Subsequent analyses will investigate how the model's predictive performance changes when image features are computed taking into consideration individual subject modifications.

Data were split two ways: First, by print pairs tested in Experiment 1 and those that were new to this experiment. Second, by whether the pairs were rated as inconclusive by at least one examiner. Model predictions are shown in Figures 9 and 10 respectively. While many of the pairs used in Experiment 1 had qualitatively good accuracy predictions, six had observed performances that were drastically different from predicted performance. Many more new pairs had inaccurate predictions. However, out of all of the pairs that were presented in Experiment 1 and had inaccurate predictions, only one had no examiners rate it as inconclusive (predicted accuracy: 79.3%, observed accuracy: 100%, number of examiners: 10).

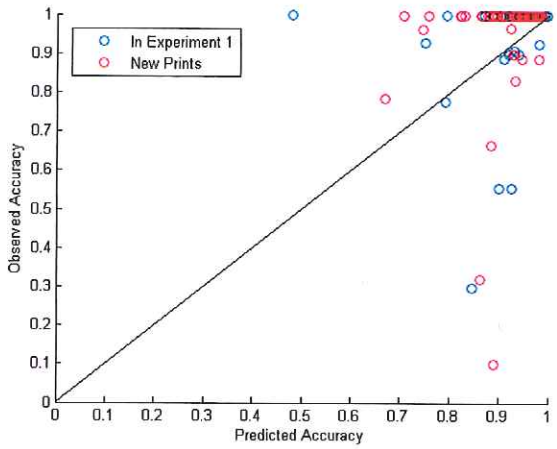


Figure 9. Predicted vs. observed accuracy for print pairs in Experiment 3. Predicted accuracy is obtained by fitting the model from Experiment 1. Pairs are split by whether they were included in Experiment 1 (blue circles) or not (red circles). There were 60 pairs in each group.

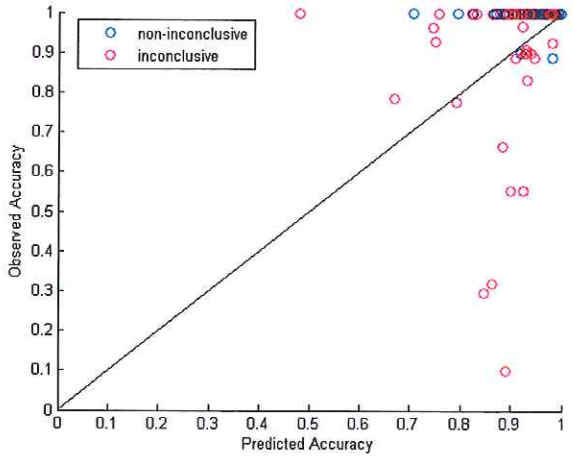


Figure 10. Predicted vs. observed accuracy for print pairs in Experiment 3 as in Figure 8. Pairs are split by whether at least one examiner rated that pair as inconclusive (red circles) or whether no examiner rated them as inconclusive (blue circles).

IV. Conclusions

In Experiment 1, we evaluated expert performance on a fingerprint matching task. Experts were highly accurate, committing few errors despite limited access to resources and restricted viewing time. Using a number of potential predictors derived from image processing algorithms, we were able to identify, using regression analyses, several image characteristics predictive of expert performance. Six features in particular were found to be important predictors of accuracy: Ridge Sum, Area Ratio, visibility of Deltas in the latent print, Mean Block Contrast of the known print, interaction between SD Block Contrast for latents and known prints, and the interaction between DEAI (deviation from expected average intensity) for the latents and known prints. Taken together, these features can explain 64% of the variance in performance accuracy on a novel set of print pairs that were withheld from those used to train the model. A classifier derived from the full data set identified the pairs on which at least one expert made a mistake with 91% accuracy, and a similar model derived from 90% of the data classified novel pairs with 75% accuracy.

Many of the same image characteristics were also predictors of subjective difficulty ratings, confidence ratings, and response times. We also found that difficulty ratings, a subjective measure, were moderately correlated with accuracy and could improve the performance of the classifier on novel print pairs.

There are several interesting observations that can be made about the set of features that we found to be predictive of accuracy (Table 2). First, four of the six features were relational, in the sense that they were calculated based on information from both prints in a comparison pair. This is a desirable feature of the model since a particular print could arise in two separate comparisons (e.g., a latent print compared to a matching and a non-matching print). In real world scenarios, a single latent print may be compared to a many known prints. In cases where one of the prints in the study was of very poor quality, such relational features might not matter. For example, if Mean Block Contrast (K) is low (i.e., for a very washed out or very dark print), then a comparison would be difficult irrespective of some relational features such as Area Ratio. Conversely, if two prints do not share Level 1 pattern type, they will not make for a difficult comparison regardless of the quality and quantity of information in each. In general, however, error rates and difficulty seem likely to be primarily characteristics of print comparisons, rather than individual prints, as difficulty for actual non-match comparisons will be most acute when the prints share significant similarities, and difficulty for actual matches will be most acute when latent quality or quantity is limited or misleading. Results in our regression models support this idea.

Second, the features within the model correspond to many types of information content. Mean Block Contrast (K), SD Block Contrast (L x K), and DEAI (L x K) capture properties of the image itself (i.e., dark or light, uniform or not). Area Ratio and Delta (L) reflect large-scale or configural (Level I) characteristics of prints, and Ridge Sum relates to visibility of fine detail in the image such as Level II features (see Introduction). These outcomes fit broadly with the idea that fingerprint examiners access different kinds of information in making comparisons and that basic image characteristics determine the detectability of relevant features and patterns.

Third, although not our primary result, the signs of the coefficients provide appealing interpretations and verify our expectations about the negative impact of low quality prints. That high contrast and clarity of ridges are predictors of accuracy should not be surprising. The DEAI measure increases as the average pixel intensity approaches 127.5, the mean expected pixel intensity for an image that contains 50% white and 50% black pixels. We assumed that this proportion would correspond to greater clarity, since a mostly light or dark image could be difficult to analyze. The positive coefficient found for this measure in the accuracy model indicates that as the proportion of white to black pixels approaches 0.5 in the latent and known print, accuracy increases. Visibility of deltas in the latent image also had a positive effect on accuracy perhaps because they provided orienting information, making it easier to match relative locations on the latent and known print. Accuracy decreased as SD Block Contrast (L x K) and Area Ratio increased. When SD Block Contrast is high in both the latent and known print, accuracy is low. In general, high variability in Block Contrast picks up variable image quality across image regions (e.g., due to gaps or smudging in portions of a print). In smudged regions, pixels would be uniformly dark, while in clear regions pixel intensity would be more variable, leading to higher contrast measures in those areas. If an image were more uniform in pixel intensities, it would have lower variability in contrast across regions and therefore lower SD Block Contrast measures. Area Ratio had a large, negative coefficient. This at first seems counterintuitive; higher area ratios tend to correspond to larger areas of latent prints. Larger areas, however, may make comparisons more difficult by making it more difficult to identify distinctive regions of the image. Since non-matching known prints were chosen by submitting the latents to an AFIS system, the non-matches likely shared many features. If experts were only shown a small latent region, it might have been easier to compare that region to the corresponding region on the known

print and quickly exclude mismatched pairs as compared to a larger latent image with more accidentally matching regions.

In addition to being able to predict accuracy, it may be important to identify which comparisons are likely to yield an error and therefore may require more scrutiny. To address this issue, we created a classifier that sorted the print pairs into ones that had perfect accuracy (so-called "perfect pairs") and ones on which at least one expert made a mistake ("non-perfect pairs"). The classifier was able to correctly sort the pairs with 91% accuracy on the training set and 75% accuracy on the testing set.

Difficulty ratings were used in two ways to add to the modeling results. We used difficulty rating itself as a predictor of accuracy. Difficulty ratings improved the fit of a model trained on all of the print pairs, but did not improve the predictive power of a model on a testing set of withheld prints. Classification performance, however, was improved. While ratings are not objective, there was nevertheless a moderate correlation between them and accuracy, suggesting that experts were aware of which comparisons were difficult, an issue we are also exploring in a paper in progress. Outside the experimental setting, it may be impractical to expect to be able to get a group of experts to provide ratings.

Difficulty ratings, confidence ratings, and response times were also evaluated as separate dependent measures. Because these measures correlated moderately with accuracy, we expected that similar features should be selected for when the same features were used to predict other dependent measures. Four of the six features that were significant predictors in the accuracy model were also significant predictors in the other models. A fifth feature, SD Block Contrast (K), which was included as part of an interaction term in the accuracy model was also included in the model of response time. Some features, such as visibility of cores, were significant predictors in the other models but not in the model of accuracy. Cores and deltas are global features. Their presence or absence can be used as a quick measure of assessing difficulty. However, global features on their own are not sufficient to make a comparison. Accuracy, therefore, depends to a greater extent on image quality, relational information, and ridge information.

These results suggest that physical characteristics of fingerprints, measured using automated image processing methods, may be valuable in predicting comparison difficulty and error rates for print pairs. Given that the present work is the first effort we know of to systematically predict errors from physical characteristics of print pairs, the predictive results are highly encouraging. Validation across larger data sets would be desirable for practical use of a predictive model such as the one derived here. Further developments along these lines, along with continuing progress in characterizing the physical quality of prints (e.g., Pulsifer et al., 2013), will likely prove to have practical value in quantifying the likely evidentiary value of expert assessments of fingerprint matches.

While these results on modeling print-pair difficulty are encouraging, there are also many differences between the paradigm used in the present study and the actual process of fingerprint comparison. Experts typically have unlimited evaluation time and access to image processing tools that were not available in the experiment described here. In addition, examiners typically are not in a "forced-choice" situation, and may decide that a real-world comparison is inconclusive. (Experiment 3, however, does attempt to correct in part for these limitations.)

Despite these limitations, there are several important dimensions to these results. The results show that even under constraints, experts were highly accurate. More than half of the print pairs had perfect accuracy, even in circumstances where the examiners' time was limited, their access to tools constrained, and they were not permitted to select the option of "inconclusive". Relatively few

studies have examined expert performance in fingerprint matching tasks, and this study adds to that body of research. It is possible, however, that error rates in forensic laboratory settings are lower than those we observed. It is also possible that other aspects of real world settings – like the danger of cognitive bias, the pressure of casework, and knowledge of extraneous information about the case – could elevate error rates as compared to experimental conditions. Care must be taken not to generalize too quickly from experimental settings, but nonetheless, such experiments can reveal a great deal about examiner performance, albeit under constraints.

Experiments in ecologically valid settings are difficult to conduct. In such settings, there are many factors that may improve accuracy (such as more time to conduct the comparisons, verification checks, etc.), as well as factors that can reduce accuracy (such as biasing influences from extraneous contextual case information, see Kassin, Dror, & Kukucka, 2013; Dror & Rosenthal, 2008). Given the significant differences between our experimental conditions and ecologically-valid fingerprint identification, we wish to reiterate that the point of the experiment reported here is not to measure such error rates, and it would be a mistake to take these data as direct evidence of a specific error rate for the field (Koehler, 2008). Rather, we are interested in identifying the features that correlate with difficulty, in order both to understand what features of print pairs affect difficulty, and to begin to understand how error rate might *vary* with comparison difficulty.

Consistent with several previous studies, very large performance differences were observed in Experiment 2, between experts and novices. Experts committed relatively few errors (approximately 9%), while novices performed nearly at chance. Experts outperformed novices despite that fact that they were under time constraints and did not have access to typical tools (i.e., image manipulation software, compass, or magnifying lens). Novices who watched a brief training video prior to the task did not perform differently overall (54% accuracy); however, trained novices committed fewer false alarms than untrained novices and, in general, were more conservative in their responses. In this manner, they were, to a limited degree, in between untrained novices and experts in at least one respect: untrained novices had many more correct answers when the prints actually matched (hit) while experts had more correct answers when the prints were from different sources (correct rejections). Trained novices performed like neither of these other groups, in that they had similar performance for both kinds of comparisons. This may reflect a shift in bias regarding an implicit 'default' conclusion – when novices see two prints with a lot of information, they may be biased to say that they match, being at a loss of what parts of the image are relevant for comparison. Experts, on the other hand, have a better sense of what features are important for making comparisons and also may be biased against false alarms (which in real world settings would result in a false conviction), saying that two prints do not match when they are unsure and therefore leading to more correct rejections. This possibility was reflected in higher confidence ratings by experts for comparisons of matching prints than for non-matching prints. Trained novices may have picked up, even on the basis of a very short video training, some idea of what information to focus on in the print and so become less likely to say that two prints match when they are unsure. Furthermore, several trained novices greatly outperformed untrained novices, with one having an overall accuracy of 75%, while the highest untrained novice accuracy was 62%.

There were also marked differences in confidence and difficulty ratings between both groups of novices and experts. In general, experts were more likely to rate prints as easy and to have higher confidence in their ratings. The short training video did not have an effect on confidence ratings among novices, but trained novices did rate comparisons as more difficult overall than untrained novices. This confirms the notion that novices were guessing when it came to comparisons, which is why their accuracy was at chance. It was not surprising that the short five minutes training video did not drastically improve performance. What was surprising is that even such a short training session

made the subjects more attuned to the difficulties of comparisons, perhaps by directing their attention to relevant features so that they became more aware of the difficulty of the task.

The same model-fitting procedure as in Experiment 1 was used to fit accuracy data for trained and untrained novices. Ridge Sum was the only predictor that appeared in all three models. Area Ratio was only selected for in the expert model. Mean Block Contrast (K), SD Block Contrast (LxK) and DEAI (LxK) appeared in models for both experts and untrained novices, but with opposite signs. This may mean that novices did not use the information appropriately. Features of images that normally help experts may have served to confuse novices. An overabundance of information may overwhelm a novice observer and lead them to incorrectly treat two complex visual stimuli as sufficiently similar. The fact that the predictors are not selected for in the model for trained novices may indicate that the training helped novices use the information in fingerprint images more appropriately. High information content did not bias them in the same way to label a comparison as a match. However, the model fits were much worse for both trained and untrained novices compared to experts, suggesting that the model did not provide a good fit to the data. This is not surprising given that accuracy was at chance for both groups.

Overall, we confirm that novices are very poor at fingerprint comparison, at least when tested on reasonably difficult exemplars. Similar to Tangen et al. (2011), we found that untrained novices had better performance for matches than non-matches. However, their match performance was not as high as that observed by Tangen et al., (62% vs. 75%). Averaged across match and non-match comparisons, novices in Experiment 2 performed at chance. Watching a short training video eliminated the difference in performance between matches and non-matches and slightly shifted bias for a subset of the subjects. This suggests that the advantage for matches for novices is due to a biased preference to label a comparison as a match. It is interesting to note that this pattern is reversed for experts. Experts have greater accuracy for non-matches than for matches. An opposing bias may exist for experts because they are more aware of the high cost of making an incorrect identification and would prefer to err on the side of caution; even under experimental conditions that instruct them to make their best guess, they may not view a false positive and a false negative as equivalent errors. Since the bias disappeared for trained novices, the training video may have emphasized the importance of correct identification, the difficulty of comparisons, and the high cost of errors. As a result, trained novices may have been more reticent to call a comparison a match by default. The first and most rapid effect of training may therefore be to alert the observer to structures in a fingerprint image that can be used to discriminate two images. As with other perceptual learning domains, more exposure is required to learn to exploit fingerprint information content to make comparisons. This demonstrates that fingerprint examiner expertise is a perceptual learning domain and is therefore likely amenable to the same kinds of training methods that have been used in mathematics and category learning (e.g., Mettler & Kellman, *in press*; Thai, Mettler, & Kellman, 2011).

Further studies need to be conducted to trace the effects of perceptual learning on accuracy and bias. A long-term study that tests examiners through various stages of their training might be able to identify gradual changes in accuracy. Changes in accuracy may correspond to a gradual reweighting of predictor variables. Examiners just beginning their training may give weights to image features in a manner similar to novices. As training progresses, a gradual shift of which variables matter most for accuracy may occur until weights match those of examiners in Experiment 1. It would be valuable and interesting further research to examine how quickly these shifts occur and how different kinds of training might affect them.

Experiment 3 sought to extend the findings of Experiment 1 by testing examiners within substantially more realistic settings for fingerprint comparison. Examiners were given unlimited time

to make their comparisons and were provided with an array of image processing tools similar to those they would normally have access to in the course of their work. In addition, experts were given the opportunity, after giving a conclusion, to label a print pair as "inconclusive", an option they were not given in Experiment 1. However, even when a pair was labeled inconclusive, experts were still required to provide a "match" or "non-match" judgment, which was meant to reflect their best guess. In this sense, our protocol was quite different from laboratory practice, in which an 'inconclusive' determination means that the examiner does not offer any further speculation about whether the print pair does or does not share a common source. But this approach gave us important clues about the relationship between performance and an indication that a print pair lacked the quality to warrant evaluation, both for an individual examiner, and in aggregate.

Error rates in Experiment 3 were similar to those in Experiment 1 and those reported in other studies (e.g., Tangen et al., 2011). This is a valuable finding, because it suggests that error rates in the first experiment cannot therefore be solely attributed to lack of resources or time to perform comparisons. There was wide variability in the number and types of tools used by experts. Tool use was often, but not always, correlated with greater difficulty and worse performance. Intuitively, this may have occurred because more difficult comparisons necessitated additional image manipulations, but the use of manipulations was not associated with greatly improved accuracy.

There was very little agreement on which comparisons were inconclusive or not. One possible reason for this discrepancy is variation in expertise among examiners. Another reason could be variation in decision criteria – some examiners may be more willing to label a print as a match or non-match rather than inconclusive than others. If differences are due to decision criteria, then one may be able to determine objectively whether there is in fact enough information to make an identification. For example, if a model predicts very high accuracy for a particular comparison, then this may be used to encourage examiners to spend extra time evaluating a comparison before determining that there is insufficient information to make a match / non-match decision. That is, it may be possible to objectively determine whether there is or is not sufficient information in a particular print pair. This would allow one to judge whether a determination of inconclusive is correct or not. We are still actively exploring how to incorporate inconclusive judgments into the model and how they relate to measures of accuracy and performance.

The model fit in Experiment 1 was used to generate predictions for comparison accuracy in Experiment 3. While the model was successful in predicting the accuracy for many comparisons (see Figures 9 and 10), there were several comparisons for which the model made poor predictions. A close examination of those comparisons revealed that all but one of them were marked as inconclusive by at least one examiner in Experiment 3.

There are several alternative ways of analyzing the data that are still under investigation. First, as mentioned earlier, features could be recomputed based on the final settings instead of using initial values. For example, if contrast was manipulated, it may be more appropriate to use the final contrast setting since this reflects the status of the image at which the identification was made. Using the final settings would mean that the tested images may not directly correspond to those used in Experiment 1, since the performance predictions of the model defined in Experiment 1 were based on the original image settings. This would result in new model predictions for a majority of the tested prints. However, if only one subject made a particular contrast adjustment then there would only be that single evaluation from which accuracy is computed. This would make it difficult to know what true average accuracy would be for a large group of experts and one reason why we did not begin with this analysis. It is interesting to note that perhaps the sequence of image manipulations might collectively be informative for predicting accuracy. For example, seeing the

same image at several contrast settings may improve performance compared to seeing an image at just one setting.

Second, instead of using the predictor weights from Experiment 1, a new model could be fit to these data. The weights may be different due to the addition of manipulation features and added evaluation time. For example, if the ability to mark *minutiae* or the number of *minutiae* marked was a very important feature in determining accuracy, the relative importance of the other predictor variables may have been degraded. Similarly, Area Ratio may matter less when more time is provided to compare two images; with little time, a smaller latent area might focus examiner attention in a way that larger areas would not. Given unlimited time, however, regardless of whether the latent area was small or not relative to the known print area, examiners could have compared sections of it at leisure. We would expect to find that many of the same predictors that were important predictors of accuracy in Experiment 1 continue to be so for this experiment. This would confirm that the originally identified image features are indeed relevant for fingerprint identification. How much those features matter, relative to one another, might depend on the exact manner in which the comparison task is set up.

Finally, the manipulations in Experiment 3 might be used to generate new features that reflect examiner behavior. Number and relative spacing of marked *minutiae*, degrees of image rotation, number of image contrast or brightness adjustment steps, or number of levels of zoom might interact with the original set of image features. For example, when Ridge Sum is low (clarity of ridges is poor), marking *minutiae* may correlate with improved accuracy, but may not matter when Ridge Sum is high. It is important to emphasize that such features are not properties of the image themselves, but decisions made by examiners. They cannot therefore be used alone to determine the difficulty of a print, but they may be informative about what kinds of behaviors and procedures are most beneficial to generating a correct identification.

In addition, we may be able to offer insight on the relationship between 'inconclusive' determinations and performance, as well as the relationship between examiners' subjective perceptions of difficulty and their objective performance. We are engaged in further analysis on both of these questions as well.

Policy Implications and Future Research

Experiment 1 was an important step in "unpacking" error rates and their relationship to difficulty, an endeavor that has great importance to forensic science and the legal system. The mere fact that some fingerprint comparisons are highly accurate whereas others are prone to error has a wide range of implications. First, it demonstrates that error rates are indeed a function of comparison difficulty (as well as other factors), and it is therefore very limited (and can even be misleading) to talk about an overall "error rate" for the field as a whole. In this study, more than half the prints were evaluated with perfect accuracy by examiners, while one print was misclassified by 91 percent of those examiners evaluating it. Numerous others were also misclassified by multiple examiners. This experiment provides strong evidence that prints do vary in difficulty and that these variations also affect the likelihood of error. Even though it was a logical assumption that print comparisons would have this quality, establishing this point empirically has significant value. Second, this study lays down a foundation for finding objective print characteristics that can quantify the difficulty of a comparison. The model we offer provides both evidence for what specific visual criteria seem to affect difficulty, as well as a model for combining these criteria to best predict accuracy. This model illustrates the benefits of creating objective measures of difficulty for print pairs, which could be substantially more efficient and consistent than more subjective approaches to assessing difficulty. It also lays the groundwork for further study that can examine the relationship between examiners'

subjective assessments of difficulty (Neumann, et al, 2013) and a more objective approach to measuring difficulty of comparisons.

Experiment 2 confirmed the differences found by prior research between novices and experts. Novice performance was essentially at chance and we obtained similar measures to those found by Tangen et al. (2011). Interestingly, we found that even exposure to a short training video seems to alter the way that novices approach the assessment task (though it did not significantly alter their overall accuracy rate). We also found that the image features used by novices were different than those used by experts. This suggests that fingerprint expertise is a perceptual learning process that results in the improved detection of structure and relevant information in fingerprint images. Consequently, procedures that promote perceptual learning (such as sequencing techniques during training; Mettler & Kellman, 2014), may be leveraged to improve training efficiency for fingerprint examiners.

Experiment 3 demonstrated that error rates in more realistic environments were generally comparable to those in Experiment 1. This is a critical finding because it means that valuable experiments with fingerprint examiners can potentially be conducted rapidly, in controlled environments without needing to rigorously replicate the environmental settings in which identifications are normally made. This can save a great deal of time, effort, and money for future research. While realistic, rigorous examination methods are of course preferred in evaluating expertise, one may also be able to generate smaller, less realistic, but similarly accurate testing materials for use during training, for example in creating an online training curriculum. The relative consistency of results between Experiment 1 and Experiment 3 suggests that while greater ecological validity is always to be preferred, valuable information may be acquired through experiments with design constraints as well. Experiment 3 also revealed two additional important findings: (1) a lack of consistency among examiners about which prints were seen to be "inconclusive" and (2) poorer aggregate performance on prints rated "inconclusive" by anyone. This raises interesting questions for further research, as well as important policy questions about where the line between 'inconclusive' and a match conclusion should be drawn.

Consider: Of the 42/120 pairs that at least one examiner labeled inconclusive, five had 50% or more of examiners agree that they were inconclusive, with an average conclusion accuracy of 59.78%. It would seem relatively clear that if we could identify these comparisons in advance, via difficulty ratings, these would be comparisons that ought not to be assessed by examiners at all, given the susbstantial risk of error and the aggregate performance only modestly above chance. But the remaining 37 comparisons that some examiner(s) labeled inconclusive had an average accuracy of 90.19%. That is, to be sure, still a substantially higher error rate than that achieved for the prints no one deemed inconclusive, but it is also quite a high accuracy rate compared to many human endeaors. Would the better practice be for these prints, could they be identified in advance by their visual metrics, not to be assessed or no conclusion offered? Or is a roughly 10 percent chance of error low enough that we would rather have this information than otherwise? Or would it, perhaps, be best to design some special, distinct examination process for this category of prints, to gain the benefits of examiners' best judgments, while recognizing that their high degree of difficulty makes them unusually error-prone? We are still in the process of assessing the relationship between objective visual characteristics and examiner's 'inconclusive' determinations, but this example illustrates how and why objective metrics (either alone, or combined with subjective measurements by examiners) may help the design of appropriate laboratory protocols and more data-driven approaches to the field and its use of information.

Overall, a more sophisticated understanding of the relationship between error rate and difficulty should also be extremely important for the courts in weighing fingerprint evidence (and has been

highlighted by the NAS (2009) inquiry into forensic science). Courts are instructed, when assessing expert evidence, to focus on the "task at hand", and this research helps to show that fingerprint examination may vary in difficulty in ways that may be relevant to its evaluation as evidence (Daubert vs. Merrell Dow Pharmaceuticals, 1993; Kumho Tire Co. vs. Carmichael, 1999). More nuanced assessments of fingerprint task difficulty might, for example, affect how a judge understands admissibility of that specific conclusion, or what degree of certainty the expert will be allowed to express, or it might impact the weight given to a specific match conclusion by the fact-finder (Faigman, Blumenthal, Cheng, Mnookin, Murphy & Sanders, 2012). It is possible, for example, that if we could accurately identify the most difficult comparisons, they could be made use of for investigative purposes but not used as evidence in the courtroom. In this way, it is possible that many prints which currently are deemed 'inconclusive' -- and may indeed be difficult enough that they are significantly more prone to error -- could be used to provide valuable, even if more error-prone, information that could assist investigations, rather than have their analysis entirely forgone.

While our model requires further testing, it is possible that it could be piloted in such a way. To be sure, our model does not yet offer the granularity to, say, associate error rates with a set number of distinct levels of difficulty, it could be adapted to examine comparisons and to predict whether they have an unusually high or low difficulty level. The implications of these findings thus have relevance both to the court and more broadly, in that they provide vital insights that can considerably enhance the procedures used in forensic laboratories. Similar to procedures for medical triage, the need for different procedures and checks can be made to fit the difficulty of a comparison.

The understanding of what makes some comparisons more difficult than others also has implications for the selection and training of fingerprint examiners. During selection, benchmarks and skill sets could be set as criteria to ensure candidates have acquired the necessary cognitive abilities needed to perform their job adequately. In addition, in evaluating the significance of errors for trainees, better information about difficulty level will be of great assistance. Trainees who make mistakes on simpler stimuli can be distinguished from those whose errors occur only on more difficult materials; for evaluating performance, all errors are not – and should not be treated as – equal.

While further research is clearly necessary to build on these results, this research therefore provides significant steps forward for helping to establish that error rates are related to difficulty; for beginning to provide validated evidence for what visual dimensions of fingerprint comparison pairs are associated with difficulty; and for helping to tease out both examiner's metacognitive abilities and the substantial degree of examiner expertise in this domain.

Acknowledgments

Portions of this report have been excerpted or adapted from the publications and papers in preparation listed in section VI Dissemination of Research Findings. We would like to thank Clara Sao, Jenny Chun, and Rachel Older for help in the data collection, and Joe Doherty, Matt Thompson, and Jaclyn Seelagy, for helpful comments and suggestions about data analysis and interpretation. We are also grateful to Gerry Laporte at the NIJ for helpful guidance, as well as UCLA for indirect support of the project. We would also like to thank the two anonymous reviewers who provided helpful evaluations and comments on this report. And finally, a great thank you to all the participants who volunteered to take part in the construction of the database and the experimental task.

V. References

Akaike, H (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.) *Second International Symposium on Information Theory*, pp. 267-281. Budapest, Hungary: Akadémiai Kiadó.

Agresti, A (1996). *An introduction to categorical data analysis.* New Jersey: John Wiley & Sons Inc.

Agresti, A (2002). *Categorical data analysis.* New Jersey: John Wiley & Sons Inc.

Ashbaugh DR (1999). *Quantitative-qualitative friction ridge analysis: An introduction to basic and advanced ridgeology.* Florida: CRC Press.

Ashworth, ARS, & Dror, IE (2000). Object identification as a function of discriminability and learning presentations: the effect of stimulus similarity and canonical frame alignment on aircraft identification. *Journal of Experimental Psychology: Applied, 6*(2), 148–157.

Baayen, RH, Davidson, DJ, & Bates, DM (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Bates, D, Maechler, M, & Bolker, B (2012). Lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://CRAN.R-project.org/package=lme4

Booth, GD, Niccolucci, MJ, & Schuster, EG (1994). Identifying proxy sets in multiple linear regression: an aid to better coefficient interpretation. Research paper INT-470. United States Department of Agriculture, Forest Service, Ogden USA.

Breslow, NE, & Clayton, DG (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Society, 88*(421), 9-25.

Bryan, WL, & Harter, N (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review, 6*, 345-375.

Burnham, KP, & Anderson, DR (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* Second edition. New York: Springer-Verlag.

Busey, TA, Schneider, B, & Wyatte, D (2008). Expertise and the width of the visual filter in fingerprint examiners. Poster presented at the 8th Annual Meeting of the Vision Sciences Society, Naples, FL.

Busey, TA, & Vanderkolk, JR (2005). Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Research, 45*, 431-448.

Busey, TA, Yu, C, Wyatte, D, Vanderkolk, J, Parada, F, & Akavipat, R (2011). Consistency and variability among latent print examiners as revealed by eye tracking methodologies. *Journal of Forensic Identification, 61*(1), 60-91.

Chatterjee, S, & Price B (1991) *Regression analysis by example.* Second Edition. New York: John Wiley & Sons.

Charleton, D, Fraser-Mackenzie, PAF, & Dror, IE (2010). Emotional Experiences and Motivating Factors Associated with Fingerprint Analysis. *Journal of Forensic Science*, 55 (2), 385-393.

Cole, SA (2002). *Suspect Identities.* Cambridge, MA: Harvard University Press.

Cole, SA (2005). Is fingerprint identification valid? Rhetorics of reliability in fingerprint proponents' discourse. *Law & Policy, 28*(1), 109-135.

Daubert v. Merrell Dow Pharmaceuticals 509 US 579 (1993).

Dixon, P (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*, 447-456.

Dror, IE (2009). How can Francis Bacon help forensic science? The Four idols of human biases. *Jurimetrics*, 50(1), 93-110.

Dror, IE, & Charlton, D (2006). Why experts make errors. *Journal of Forensic Identification, 56*(4), 600-616.

Dror, IE, Charlton, D, & Péron, AE (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International, 156*(1), 74-78.

Dror, IE, & Cole, SA (2010). The vision in "blind" justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review, 17*(2), 161-167.

Dror, IE, Péron, AE, Hind, S & Charlton, D (2005). When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology, 19*(6), 799–809.

Dror, IE, & Mnookin, JL (2010). The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. *Law, Probability, & Risk, 9*(1), 47-67.

Dror, IE, & Rosenthal, R (2008). Meta-analytically quantifying the reliability and bias ability of forensic experts. *Journal of Forensic Sciences, 53*(4), 900-903.

Dror, IE, Stevenage, SV, & Ashworth, A (2008). Helping the cognitive system learn: Exaggerating distinctiveness and uniqueness. *Applied Cognitive Psychology, 22*(4), 573-584.

Expert Working Group on Human Factors in Latent Print Analysis (2012) Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach; NIST Interagency Report 7842.

Faigman, DL, Bllumenthal, JA, Cheng, EK, Mnookin, JL, Murphy, EE, & Sander, J (2012). Modern Scientific Evidence: The Law and Science of Expert Testimony: Fingerprints, Vol. 5 (Ch. 33)

Gelman, A, & Hill, J (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Gelman, A, & Su, Y-S (2013). arm: Data analysis using regression and multilevel/hierarchical models. R package version 1.6-04. http://CRAN.R-project.org/package=arm

Gibson, EJ (1969). *Principles of perceptual learning and development.* New York: Prentice Hall.

Green, DM, & Swets, JA (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

Haber, L & Haber, RN (2008). Scientific Validation of Fingerprint Evidence under Daubert. *Law, Probability and Risk, 7*, 87–109.

Hall, LJ, & Player, E (2008). Will the introduction of an emotional context affect fingerprint analysis and decision-making? *Forensic Science International, 181*(1-3), 36-39.

"How to Compare Fingerprints – The Basics" http://www.youtube.com/watch?v=IrpTqKkgygA

Jaeger, TF (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446.

Kang, J, Bennet, JM, Carbado, D, Casey, P, Dasgupta, N, Faigman, D, Godsil, R, Greenwald, AG, Levinson, J, & Mnookin, J (2012). Implicit Bias in the Courtroom, *UCLA Law Review*, 59, 1124-1185.

Kassin, SM, Dror, IE, & Kukucka, J (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition, 2*(1), 42-52.

Kellman, PJ (2002). Perceptual learning. In: Pashler H, Gallistel CR, editors. *Stevens' handbook of experimental psychology*, vol. 3 (3rd edition). New York: John Wiley & Sons. pp. 259-299.

Kellman, PJ, & Garrigan, P (2009). Perceptual learning and human expertise. *Physics of Life Reviews, 6*(2), 53-84.

Kellman PJ, & Massey CM (2013) Perceptual Learning, Cognition, and Expertise. In: Ross BH, editor. *The psychology of learning and motivation* vol. 58. Amsterdam: Elsevier Inc. pp. 117-165.

Koehler, JJ (2008). Fingerprint error rates and proficiency tests: What they are and why they matter. *Hastings Law Journal, 59*, 1077-1100.

Kovesi, PD (2000). MATLAB and Octave functions for computer vision and image processing. Available from http://www.csse.uwa.edu.au/~pk/research/matlabfns/

Kumho Tire Co. v. Carmichael 526 US 137 (1999).

Langenburg, G (2009). A performance study of the ACE-V process: A pilot study to measure the accuracy, precision, reproducibility, repeatability, and bias ability of conclusions resulting from the ACE-V process. *Journal of Forensic Identification, 59*(2), pp. 219-257.

Langenburg, G, Champod, C, & Wertheim, P (2009). Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. *Journal Forensic Sciences*, 54, 571-582.

Maltoni, D, Maio, D, Jain, AK, & Prabhakar, S (2009) *Handbook of fingerprint recognition.* London: Springer.

Marcon, JL (2009) *The distinctiveness effect in fingerprint identification: How the role of distinctiveness, information loss, and informational bias influence fingerprint identification* (Doctoral dissertation). Available from *ETD Collection for University of Texas, El Paso (*Paper AAI3358893).

Mettler, E, & Kellman PJ (*In Press*). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research.*

Mnookin, JL (2001). Fingerprint evidence in an age of DNA profiling. *Brooklyn Law Review*, 67, 13-70.

Mnookin, JL (2008a). Of black boxes, instruments, and experts: Testing the validity of forensic evidence. *Episteme*, *5*(3), 343-358.

Mnookin, JL (2008b). The validity of latent fingerprint identification: Confessions of a fingerprinting moderate. *Law, Probability & Risk*, *7*, 127-141.

Mnookin, JL The Courts, The National Academy of Science, and the Future of Forensic Science, 75 Brooklyn Law Review 75, 1209-1276 (2010) (The Ira. M. Belfer Lecture, 2009).

National Research Council, National Academy of Sciences (2009). *Strengthening Forensic Science in the United States: A Path Forward.* Washington, DC: National Academy Press.

NIST, Latent Print Examination and Human Factors: A Systems Approach (2012).

Neter, J, Kutner, MH, Wasserman, W, & Nachtsheim, C (1996). *Applied linear statistical models.* Fourth Edition. USA: McGraw-Hill/Irwin.

Neumann, C, Champod, C, Yoo, M, Gennessay T, & Langenburg, G (2013). Improving the understanding the reliability of the concept of "sufficiency" in friction ridge examination. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=266312

Oppenheimer, DM (2008). The secret life of fluency. *Trends in Cognitive Science*, *12*(6), 237-241.

Pulsifer, DP, Muhlberger, SA, Williams, SF, Shaler, RC, Lakhtakia, A (2013). An objective fingerprint quality-grading system, *Forensic Science International*, *231*, 204-207.

Schiffer, B, & Champod, C (2007). The potential (negative) influence of observational biases at the analysis stage of fingermark individualisation. *Forensic Science International*, *167*, 116-120.

Schneider W, & Shiffrin RM (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychology Review*, *84*, 1-66.

Shen, W, & Eshera, MA (2003). Feature extraction in fingerprint images. In N. Ratha and R. Bolle (Eds.), *Automatic fingerprint recognition systems* (pp. 145-182). New York: Springer-Verlag.

Tangen, JM, Thompson, MB, & McCarthy, DJ (2011). Identifying fingerprint expertise. *Psychological Science*, *22*(8), 995-997.

Thai, K, Mettler, E, & Kellman, PJ (2011). Basic information processing effects from perceptual learning in complex, real-world domain. In L. Carlson, C. Holscher, & T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 555-560). Boston, MA: Cognitive Sciences Society.

Ulrey, BT, Hicklin, RA, Buscaglia, J, & Roberts, MA (2011). Accuracy and reliability of forensic latent fingerprint decision. *Proceedings of the National Academy of Sciences*, *108*(19), 7733-7738.

Ulery, BT, Hicklin, RA, Buscaglia, J, & Roberts, MA (2012). Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. *PLoS ONE 7*(3), e32800.

Vokey, JR, Tangen, JM, & Cole, SA (2009). On the preliminary psychophysics of fingerprint identification. *Quarterly Journal of Experimental Psychology*, *62*, 1023-1040.

Wertheim, K, Langenburg, G, & Moenssens, A (2006). A report of latent print examiner accuracy during comparison training exercises. *Journal of Forensic Identification*, 56, 55-93.

Wickens, TD (2002). *Elementary Signal Detection Theory.* Oxford University Press. New York: New York.

Zuur, AF, Ieno, EN, Walker, NJ, Saveliev, AA, & Smith, GM (2009). *Mixed effects models and extensions in ecology with R.* New York: Springer Science+Business Media, LLC.

VI. Dissemination of Research Findings

Journal Articles

Kellman, PJ, Mnookin, J, Erlikhman, G, Ghose, T, Garrigan, P, Mettler, E, Charlton, D, & Dror, I. . Forensic Comparison and Matching of Fingerprints: Using Quantitative Image Measures for Estimating Error Rates through Understanding and Predicting Difficulty, PLOS ONE, 9(5): e94617. doi:10.1371/journal.pone.0094617

Erlikhman, G, Kellman, PJ, Dror, I,, Ghose, T, Garrigan, P, Charlton, D, & Mnookin, J Experts and Novices: Using Fingerprint Image Features to Understand Performance . *In Preparation*

Mnookin, J, Dror, I., Doherty, J., Seelgacy, J., Erlikhman, G, , Garrigan, P, & Kellman, PJ, Metacognition, Recognizing Difficulty, and Expert Fingerprint Examiners. *In Preparation*

Conference Presentations focusing on grant research include:

Ghose, T., Erlikhman, G., Garrigan, P., Mnookin, J., Dror, I., Charleton, D., & Kellman, P.J. (2013). Perception, Image Processing and Fingerprint-Matching Expertise. *European Conference for Vision and Perception, August 2013*

Erlikhman, G., Ghose, T., Garrigan, P., Mnookin, J., Dror, I., Mettler, E, Charleton, D., & Kellman, P.J. (2013). Fingerprint matching expertise and its determinants. *Vision Sciences Society, May 2013*

Additional Presentations referencing research (partial):

Fingerprint Evidence and Current Research: presentation at NACDL/Cardozo Law School National Forensic Science College, June 2014

Keynote presentation on "A Cognitive Perspective on Expert Evidence and the Administration of Justice" at Project Innocence Annual Meeting, Portland, 11 April 2014.

Invited presentation on "The Human Factor in Forensic Science" at the University of Amsterdam, 11 October 2013.

Keynote presentation on "Distributed Cognition Between Humans and Technology" at Biometrics Institute Technology Showcase, 27 June 2013.

The Sir Michael Davies Keynote presentation on "Experts: The Myth of Impartiality" at the Expert Witness Institute (EWI) Annual Meeting, London, 5 June 2013.

Second Workshop on "Cognitive Factors in Making Forensic Comparisons", at the London Metrapolitan Police, 22 May 2012.

Keynote presentation on "Distributed Cognition Between Human Experts and Technology" at the Annual User's Education Conference, Bellevue, 8 May 2012.

Invited presentation on "Psychology and the Law: Cognitive failing in administering justice, and how psychology can help the criminal justice system" at the University of Seattle, WA, 6 May 2013.

Keynote presentation on "Cognitive Forensic" at the Forensics Europe Expo Conference, London, 24 April 2013.

First workshop on "Cognitive Factors in Making Forensic Comparisons", at the London Metrapolitan Police, 17 April 2012.

Invited presentation on "Perception and Judgments of Human Experts: The Role of Contextual Information" at the Department of Psychology, University College London (UCL), 12

February 2013.

Invited presentation on "Cognitive Underpinning of Expertise: Why & How Forensic and Medical Experts Make Errors, and How to Minimise Them" at the Department of Psychology, Warwick University, 7 February 2013.

Second 1-day Workshop on "Cognitive Factors in Making Forensic Comparisons", at The Netherlands Forensic Institute (NFI), 10 December 2012.

A 2-day workshop on "The Human Element and Cognition in Biometric Identification" at the Biometric Institute, 28-29 November, 2012.

Keynote presentation on "Brain Friendly Biometric Systems: Effective Distributed Cognition Between Humans and Technology" at 8th Biometrics Institute Technology Showcase & Exhibition, 27 November 2012.

Invited presentation on "Contribution of Cognitive Psychology: Reliability and Biasability of Experts in the Court Room" at the University of New South Wales, Sydney, 23 November 2012.

Invited presentation on "Cognitive Forensics: Increasing expertise in forensic science" at the Centre for Forensic Science, University of Technology, Sydney, 21 November 2012.

A 2-day workshop on "Cognitive Factors in Making Forensic Comparisons" at the Australian National Institute of Forensic Science (NIFS) and Australian New Zealand Policing Advisory Agency (ANZPAA), 19-20 November 2012.

A 2-day workshop on "Cognitive Factors in Making Forensic Comparisons" at Victoria Police, 15-16 November 2012.

Invited presentation on "Cognitive Forensics: Identifying and Mitigating Bias in Criminal Cases" at the Criminal Bar Association Autumn Conference, 3 November 2012.

A 1-day workshop on "Improving Forensic Decision Making" at the Colorado Bureau of Investigation (CBI), 17 September 2012.

Plenary presentation on "Cognitive Forensics, Expertise, the Biasing Snowball Effect, and Context Management in Forensic Investigations' at the 6th Annual European Academy of Forensic Science Conference, 23 August 2012.

Invited presentation on "Expert Evidence: The Good, the Bad, and the Ugly" at the After Court Seminar program for High Court Judges, Deputy High Court Judges, Judges of the Court of Appeal, and Justices of the Supreme Court, Royal Courts of Justice, 13 June 2012.

From:	Vanessa Antoun
To:	FN-OSTP-PCAST
Cc:	Norman Reimer
Subject:	NACDL response to PCAST request
Date:	Wednesday, December 14, 2016 5:45:57 PM

Dear PCAST,

Thank you for inviting NACDL to reply to a request for information as a follow-up to the President's Council of Advisors on Science and Technology Report to the President on "Forensic Science in the Criminal Courts: Ensuring Scientific Validity Of Feature-Comparison Methods." NACDL provides the following response regarding microscopic hair comparison: We are not aware of any appropriately designed research studies that provide sufficient empirical evidence to establish the foundational validity and estimate the accuracy of microscopic hair comparison analysis.

If you have additional questions or NACDL can be of further assistance, please let me know.

Thank you,
Vanessa Antoun

Vanessa Antoun
Senior Resource Counsel
National Association of Criminal Defense Lawyers
1660 L Street NW, 12th Floor
Washington, DC 20036
Phone: (202) 465-7663
Fax: (202) 872-8690
vantoun@nacdl.org

From:	Schwartz, Ted
To:	FN-OSTP-PCAST
Subject:	Articles on Footwear analysis
Date:	Wednesday, December 14, 2016 2:45:58 PM
Attachments:	Footwear Examinations Mathematical Probabilities of Theoretical Individual Characteristics.pdf Statistical Discrimination of Footwear.pdf The Mount Bierstadt Study An Experiment in Unique Damage Formation in Footwear.pdf the science of tire impression identification.pdf

---

Please find four (4) articles that deal with the significance of a "match" in footwear and tire comparisons.

Thank you.

Ted R. Schwartz
Senior Forensic Scientist
Westchester County Forensic Laboratory
10 Dana Rd, Valhalla, NY 10595
914-231-1630

Article

Footwear Examinations: Mathematical Probabilities of Theoretical Individual Characteristics

Rocky S. Stone

Albuquerque Metropolitan Crime Laboratory (retired)
Albuquerque, NM

Abstract: The trend in the forensic sciences favors objectivity over subjectivity. Courts in the United States are becoming increasingly hesitant to accept the opinion of an examiner who states, "It's a 'match' because I *say* it's a 'match'". Objectivity, in most cases, is reinforced by quantification. The individual characteristics that appear on a shoe print or shoe impression can be quantified using two primary variables. Their location on the print and their configuration and orientation yield measurable, discriminating data values. Theoretical types of individual characteristics that are found on shoe prints are described and discussed, and a hypothetical model is presented with probability estimates applied to quantify the likelihood of occurrence of the characteristics. With marks or combinations of marks of reasonable complexity, the magnitudes of the resultant numbers, though entirely abstract and based upon conservative assumptions, are remarkable.

Introduction

The presence of accidental, random defects on a shoe leaving a print* may allow an examiner to "positively identify" that particular shoe to the exclusion of all other shoes as having created the print. The assumed underlying premise is that nature never repeats itself. When physical entities, both natural and man-made, are examined in sufficiently fine detail, the

* Throughout this discussion, unless specifically differentiated, the term "print" will be used to refer to both three-dimensional shoe impressions and two-dimensional prints.

Received January 3, 2006; accepted February 27, 2006

JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, January 2010, Vol. 55, No. 1
doi: 10.1111/j.1556-4029.2009.01209.x
Available online at: interscience.wiley.com

PAPER

CRIMINALISTICS

Nicholas D. K. Petraco,[1,2] *Ph.D.; Carol Gambino,*[1] *M.S.; Thomas A. Kubic,*[1,2,3] *Ph.D.; Dayhana Olivio,*[1]*; and Nicholas Petraco,*[1,3,4] *M.S.*

Statistical Discrimination of Footwear: A Method for the Comparison of Accidentals on Shoe Outsoles Inspired by Facial Recognition Techniques

ABSTRACT: In the field of forensic footwear examination, it is a widely held belief that patterns of accidental marks found on footwear and footwear impressions possess a high degree of "uniqueness." This belief, however, has not been thoroughly studied in a numerical way using controlled experiments. As a result, this form of valuable physical evidence has been the subject of admissibility challenges. In this study, we apply statistical techniques used in facial pattern recognition, to a minimal set of information gleaned from accidental patterns. That is, in order to maximize the amount of potential similarity between patterns, we only use the coordinate locations of accidental marks (on the top portion of a footwear impression) to characterize the entire pattern. This allows us to numerically gauge how similar two patterns are to one another in a worst-case scenario, i.e., in the absence of a tremendous amount of information normally available to the footwear examiner such as accidental mark size and shape. The patterns were recorded from the top portion of the shoe soles (i.e., not the heel) of five shoe pairs. All shoes were the same make and model and all were worn by the same person for a period of 30 days. We found that in 20–30 dimensional principal component (PC) space (99.5% variance retained), patterns from the same shoe, even at different points in time, tended to cluster closer to each other than patterns from different shoes. Correct shoe identification rates using maximum likelihood linear classification analysis and the hold-one-out procedure ranged from 81% to 100%. Although low in variance, three-dimensional PC plots were made and generally corroborated the findings in the much higher dimensional PC-space. This study is intended to be a starting point for future research to build statistical models on the formation and evolution of accidental patterns.

KEYWORDS: forensic science, footwear, shoes, multivariate, principal component analysis, linear discriminant analysis, pattern recognition, accidental marks, accidentals

Footwear impression evidence is present at many crime scenes and can be found visible or latent on a variety of surfaces such as glass, carpet, paper, wood, dirt, concrete, tile, and snow (1). Shoe impressions can be more of a challenge for a criminal to avoid leaving than fingerprints, and like fingerprints, they can link a person to a crime scene (1). Nevertheless, footwear impression evidence is much less utilized because it is more difficult to spot and collect and more prone to contamination. Also, the ability to make positive identifications between a suspect's shoes and crime scene impressions is not as well known to those in criminal law (1).

Shoe impressions can be identified based on class characteristics like manufacturer, brand, model, and shoe size (2). There are thousands of different shoe designs for men and women, as well as a variety of sizes. The rapid rate that shoe designs are replaced adds to the discriminating power of shoe print evidence. Aside from design, the possible imperfections, variations, and random characteristics introduced during the manufacturing process can significantly reduce the number of possible candidates in the identification of an unknown impression (1).

Accidental characteristics (accidentals) are nonreproducible cuts, tears, punctures, and the like that accumulate on the outsole as the shoe is worn (2). Much like minutiae on fingerprints, footwear accidentals are identified based on agreement in a feature's appearance and position. Fingerprint minutiae, however, only have a finite number of descriptors. The possible shapes of a shoe accidental mark are infinite (1,3,4). Hence, if the shape of an accidental has enough complexity, it is theorized that just one would be enough to make a positive identification (1).

While it is a strongly held belief by many footwear examiners that the patterns of accidental marks on shoes are unique, this is an inductive conclusion that has not been thoroughly studied using controlled experiments. This poses a problem in the wake of the Daubert decision in which the U.S. Supreme Court rejected the Frye "general acceptance rule" concerning the admissibility of certain evidence submitted as scientific (5,6). As a result, various forms of physical evidence have been the subject of Daubert and other admissibility challenges. While the use of footwear impression evidence in criminal trials has recently been upheld by the United States Court of Appeals in a 2006 Daubert challenge case, future challenges are inevitable (7).

Taken literally, the adjective "unique" applied to accidental patterns means that there is one and only one pattern like it in the

[1]Department of Science, John Jay College of Criminal Justice, City University of New York, 899 10th Avenue, New York, NY 10019.
[2]Faculty of Chemistry, Graduate Center, City University of New York, 365 5th Avenue, New York, NY 10016.
[3]Faculty of Forensic Science, Graduate Center, City University of New York, 365 5th Avenue, New York, NY 10016.
[4]New York City Police Department Crime Laboratory, Trace Evidence Unit, 150-14 Jamaica Avenue, Jamaica, NY 11432.

Received 6 Oct. 2007; and in revised form 12 Jan. 2009; accepted 31 Jan. 2009.

world. This is a conclusion impossible to prove unless the accidental pattern of every shoe in use in the world is known at every point in time. With such a seemingly impossible task at hand, i.e., to "prove the uniqueness" of an accidental pattern, does this mean all is lost? Of course not! There exists a vast array of statistically based methodologies which, when given complicated pattern data, can be used to gauge similarity in a statistical sense. Many of these methods are known to be robust even using small to medium sample sizes (8,9). Other methods, given enough data, should be able to yield random match probabilities (9–11). Unfortunately, few attempts have been made to apply actual mathematical formulas to the study of accidental patterns. A way to buffer against admissibility challenges (e.g., Daubert challenges) is to analyze the data with statistical methods. Once implemented, these forensic pattern comparison systems can be extensively tested and identification error rates can be established. In this way, one can generate quantitative comparisons based on sound scientific (statistical) principles and lend objectivity and reliability to a field seen largely as subjective.

Everett, Lambert, and Buckleton have suggested a Bayesian approach to interpreting footwear marks (12). They advocate applying their method to footwear identification when the acquired features alone are not overwhelming enough to warrant a sound positive identification. Another study was performed by Geradts et al. who used algorithms to construct a footwear database called REBEZO in cooperation with the Dutch police. The data consisted of shoeprints found at crime scenes, shoes obtained from suspects, and store-bought shoes (13,14). The algorithm segments shoe sole profiles and attempts to identify and classify distinguishable shapes for comparison against a database of known shoe sole profiles. The authors note that currently the system has difficulty comparing complex shapes and that more research is needed (15).

Computational models of facial recognition have proved extremely successful in criminal investigation and security systems. These numerical pattern comparison techniques are fast, reliable, and relatively easy to understand. While there are some differences from one model to the next, they all attempt to represent a facial image as a data set which is compared to other data sets stored in a database (16–18). Such data sets will obviously be very complex, and distinguishing between them requires use of a computer that can sort through vast amounts of data and perform complicated pattern recognition tasks. Using computers in pattern recognition has the added benefit of lessening human bias introduced in gauging how "similar" two patterns of data are to one another (10,11,15,18,19).

In facial pattern recognition, a particular scheme stemming from information theory decomposes the data set representation of an image (a facial image, an accidental pattern on a shoe sole, etc.) into a smaller set of characteristic "features" known as principal components (PCs) (20,21). Principal component analysis (PCA) essentially eliminates information which varies little within all the accidental patterns included in the analysis, and captures the variation within the data independently of any human judgment. The method accomplishes this task using new statistically uncorrelated and orthogonal variables constructed from the old variables. PCA serves to reduce the dimension of the data, which can be enormous, to a manageable level.

The purpose of this study is to use facial pattern recognition techniques to demonstrate that accidental patterns found on footwear outsoles can be compared against each other within a statistical pattern recognition framework. Using such a framework we then show how identification error rates of the system can be estimated. In this study, we only used the accidental pattern from the top of the sole, and only the positions of the accidentals were recorded. Size and shape were not used in comparisons due to the fact that characteristics having amorphous properties are very difficult to treat computationally (22,23). While our programs are evolving to take these features into account, as of yet, they cannot. Second, our minimal treatment of the accidental patterns examined in this study allow us to show how robust statistical discriminations can be made even with a minimal amount of information, i.e., using only the distribution of the accidentals on the top sole impression.

In this study, five pairs of shoes having the same manufacturer and model were worn by the same person for a period of 30 days each. Although there have been studies monitoring the appearance of accidental characteristics over time (2), there have been none on the same model shoe worn by the same person. Under these circumstances, one would expect the greatest positional agreement of accidentals as many of the usual variables will be constant such as manufacturer, material, foot morphology, weight distribution, walking pattern, and routine. Aside from observing how similar the footwear patterns will be under these conditions, this study will provide a starting point for future research upon which to build statistical information on the formation and evolution of accidental patterns.

Methodology

Five pairs of ladies Lands' End, size 7 med (model D86 M30400 565) shoes were worn for a period of 30 days each. An initial shoe print was recorded before wear. Shoe prints were subsequently recorded on days 1 through 7, 14, 16, 18, 20, 24, 28, and 30. A total of 15 patterns were recorded for each shoe. Four replicates were made of each print for evaluating repeatability and because the first print made was always too dark to see fine detail. The naming convention used for each shoe distinguished order of pair worn and left or right. For example, the first shoe pair worn is called P1. The left shoe of P1 is called P1L and the right shoe is called P1R.

Unfortunately most of the accidental patterns for the left shoe of pair 5 (P5L) were not readable and thus all the patterns for P5L were dropped from this study. Hence, there are nine shoes in this study: P1L (shoe 1), P1R (shoe 2), P2L (shoe 3), P2R (shoe 4), P3L (shoe 5), P3R (shoe 6), P4L (shoe 7), P4R (shoe 8), and P5R (shoe 9).

Generation of Outsole Prints

The magna brush method was used to record the prints on to 8″ × 11″ white copy paper as it is found to be superior to dusting fingerprint powders when dealing with nonsmooth and porous surfaces (24). Black magnetic flake powder (all particles magnetic) was used over magna powder (iron particles mixed with fine powder) as it enhances wet prints better, and produces much lower background and smudge levels (24).

Recording Accidental Marks

Only the position and quantity of accidentals were considered in this study. Neither their size nor morphology was evaluated. Accidentals were recorded using a charting method adapted from the Abbott grid locater and the method adapted from a paper involving statistical analysis of barefoot impressions performed by Kennedy et al. (25–27). Figure 1 shows the grid used in this study to record the accidental patterns.

Using the prints from the magna brush technique, the best replicate from each interval was selected. Two lines were drawn tangent

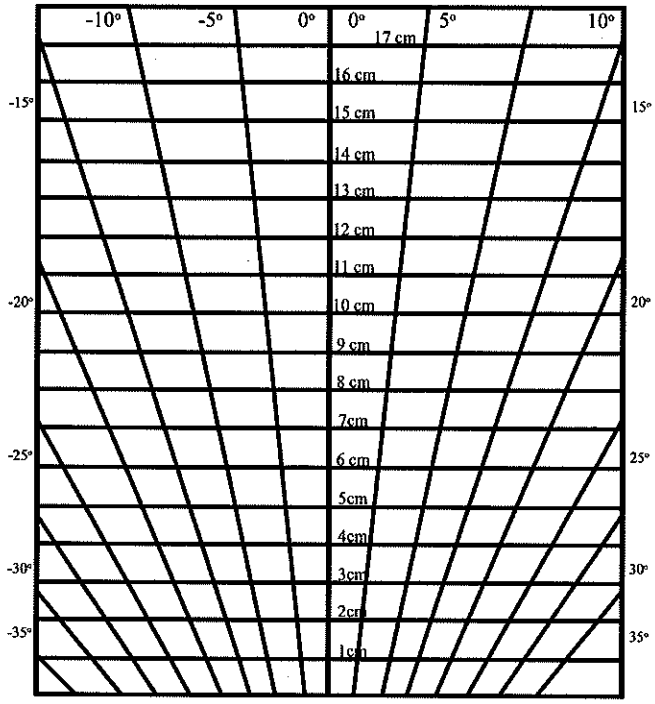


FIG. 1—*Grid used to record the accidental patterns.*

to the widest part of the print on the ball portion of the shoe's sole, and another two lines tangent to the widest part of the print on the heel of the shoe. The midpoint between each of these was marked. Another line was made perpendicular to the tangent lines, directly under where the shoe pattern of the sole ended. The two midpoints of the ball and heel were then connected.

The prepared grid (cf. Fig. 1) was printed on transparent film so that it could be laid directly over the shoe print. The X axis of the grid was placed over the perpendicular line drawn below the end of the sole's pattern. The Y axis was placed over the line that connects the two midpoints. In this way, all accidentals in the ball part of the shoe could be characterized. In this study, only the top portion of the shoe sole was examined, not the heel.

The number of accidentals found in each quadrant was recorded into charts for each shoe and day of wear. In order for an accidental to be counted it must have appeared clearly in more than one replicate. If an accidental extended into two different quadrants, the quadrant was selected in which the majority of the accidental was found. Because of wear and variations in the manufacturing process and in making each print (the amount of oil used, how one stepped on the paper, etc.), the shoeprint proportions were not exactly the same for each print. Thus, using the above method of preparing the print for the transparent grid may yield different results (the area on the sole in which the boxes lay may be different). In order to deal with this problem a shoeprint for both the right and left foot was selected to be used as a master template whose proportions were used for all other prints regardless of their own proportions. In this way, the boxes of the grid were positioned over the same place for each print.

Statistical Methods

In facial recognition, a two-dimensional image is numerically represented as a vector (a one-dimensional list) of pixels that make it up. For example, a 256×256 pixel image is rearranged into a 65,536 unit long vector. Each pixel is a box of varying color and intensity. An accidental pattern on the outsole of a shoe may also be "pixilated" or rather divided up into boxes and rearranged into a vector. Each box that makes up the accidental pattern contains a varying number of accidental marks. The accidental patterns are compared by first decomposing them into their PCs and then using a metric function to measure their distance apart in PC-space. We will use the method of maximum likelihood Gaussian linear classification analysis (sometimes also called linear discriminant analysis, LDA) to numerically gauge the similarity between patterns based on their proximity in PC-space (10,11,28).

The grid for recording the positions of accidental marks was 18×18 boxes as shown in Fig. 1. The boxes, with the number of accidentals they contained, were translated into feature vectors $\mathbf{X}_i$ (9,19). In this study, the components of the feature vector are the number of accidental marks appearing in a particular grid box of a given shoe on a given day. Data were stored in Excel and algorithms for the accidental pattern comparisons were written using the Mathematica computer algebra system (29). Initially, a total of fifteen accidental patterns were to be recorded for each of ten shoes (five pairs) for a total of 150 accidental patterns, but because accidental marks for the left shoe of the fifth pair came out unreadable, only 135 accidental patterns were recorded. Furthermore, only those accidental patterns with at least one accidental mark were included in the PCA and maximum likelihood Gaussian—linear classification analysis (MLG-LCA), leaving a total of 116 accidental patterns.

The translation of the accidental patterns into feature vectors was performed by stacking each 18-unit-long column of the grid beneath the one after it starting from the leftmost column. This procedure yielded feature vectors 324 units long (18×18 boxes = 324 boxes) which was then assembled into an $n \times P$ data matrix $\mathbf{X}$, where n is the number of accidental patterns (feature vectors, here 116) to be used in a given analysis, and $P = 324$.

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,j} & \cdots & X_{1,324} \\ \vdots & & \vdots & & \vdots \\ X_{i,1} & \cdots & X_{i,j} & \cdots & X_{i,324} \\ \vdots & & \vdots & & \vdots \\ X_{116,1} & \cdots & X_{116,j} & \cdots & X_{116,324} \end{bmatrix}$$

Every box in the feature vector represents a random variable and every row in the data matrix is a vector of number of accidental marks observed. The symbol $\mathbf{X}_i$, designates a (row) vector of data representing accidental pattern i. A data matrix with n rows contains n accidental patterns. The average of all row vectors in $\mathbf{X}$ is the average vector $\overline{\mathbf{Z}}_i$. The multivariate analyses of data set ($\mathbf{X}$) undertaken in this study were PCA and MLG-LCA. For details on these methods see reference (30) and references therein. The Mathematica notebooks developed for this study are available upon request from the authors.

Since the feature vector of an accidental pattern is simply a point (in a high dimensional space) the similarity between patterns can be gauged by an appropriate distance metric and decision algorithm. The degree of "sameness" between two arbitrary accidental patterns was determined numerically by using MLG-LCA (28). The PC-derived data matrix $\mathbf{Z}$ was used in place of the original data matrix of accidental patterns $\mathbf{X}$, due to its significantly smaller size (116×32 at most for $\mathbf{Z}$ vs.116×324 for $\mathbf{X}$ in one case) and due to the fact that direct application of MLG-LCA to $\mathbf{X}$ was impossible due to problems encountered with singular pooled covariance matrices required by the algorithm.

The MLG-LCA decision model uses a distance function to find the mean feature vector $\overline{\mathbf{Z}}_i$ that is closest to the "test" feature vector $\mathbf{Z}_j$.

The actual discriminant function constructed for the patterns from shoe i is given as

$$L_i(\mathbf{Z}_j) = \overline{\mathbf{Z}}_i^{\mathrm{T}} \mathbf{S}_{\mathrm{pl}}^{-1} \mathbf{Z}_j - \frac{1}{2} \overline{\mathbf{Z}}_i^{\mathrm{T}} \mathbf{S}_{\mathrm{pl}}^{-1} \overline{\mathbf{Z}}_i$$

where $\mathbf{S}_{\mathrm{pl}}^{-1}$ is the inverse of the pooled covariance matrix. See reference (30) for a detailed description. A total of $k = 9$ discriminant functions were constructed, one for each shoe. The algorithm decides that the accidental pattern j is most similar to the predefined set of accidental patterns from shoe i according to the decision rule

$$\arg\max_i L_i(\mathbf{Z}_j).$$

The predefined sets of accidental patterns were chosen to be all those patterns recorded for a particular shoe over the course of the 30 days. Thus, there are nine sets of accidental patterns, one set for each shoe. Each set consists of 15 patterns, one for each day that data was recorded. In words, the decision rule above means: "the (PCA reduced) accidental pattern $\mathbf{Z}_j$ is most similar to the set of (PCA reduced) accidental patterns from shoe i whose discriminant function yields the largest value" (28).

The ability of discriminant functions to accurately predict the sample identity of a pattern which they have not been trained with, is called classification error analysis (31). This is a very important topic whenever statistical pattern recognition techniques are applied to forensic evidence. This is because discriminant functions, while trained on a finite (probably small) set of data, will be expected to classify or identify new pieces of evidence which they have not been trained with. Thus, rigorously derived accurate estimates for error rates of computed sets of discriminant functions are critical in forensic science applications. For this study, we estimate the error rates of the k discriminant functions in three different ways. We actually compute estimates of the "correct classification rate" which is one minus the error rate and is reported as a percentage.

The first estimate used is the "apparent" correct classification rate computed by determining the number of accidental patterns assigned to their correct sample (by the discriminant functions) divided by the total number of accidental patterns. This performance estimate is known to be biased and tends to yield an overly optimistic correct classification rate (28).

The second estimate is the overall "hold-one-out" correct classification rate (10,32). This is computed by first recalculating the linear discriminant functions omitting a single accidental pattern from the data set. Thus, the recalculated discriminant functions are not trained to identify the held out accidental pattern. This omitted accidental pattern is then classified with the recalculated linear discriminant functions and the process is repeated sequentially for each accidental pattern in the data set. The number of correctly classified "held-out" accidental patterns is divided by the total number of accidental patterns in the entire data set (116 in this study) to yield the overall hold-one-out correct classification rate.

Finally, the "average hold-one-out" correct classification rate is computed. This process involves replicating a data set composed of n observation vectors, n-times. Each replicate data set, however, contains all but one of the original data vectors (33). The n data sets are then used to recalculate a statistic on that data set in the absence of the deleted data vector, producing a set of estimates of the statistic. The set can then be used to produce an average and standard deviation for the statistic (33). The average hold-one-out correct classification rate is mathematically the least biased estimation of the discrimination functions' classification performance (13).

Here, we compute the average hold-one-out correct classification rate by first computing all the samples' hold-one-out correct classification rates and recording them in a "cross-validation table." Next, the average and standard deviation of the samples' hold-one-out correct classification rates is found yielding the average hold-one-out correct classification rate (10,32,33).

Results and Discussion

PCA of All Accidental Patterns with At Least One Accidental Mark

Out of 135 accidental patterns recorded for nine shoes (cf. Methodology section, paragraph two), 116 contained at least one accidental mark. These 116 accidental patterns were processed with PCA. It was found that for the shoes in this study (all worn by the same person and all the same make and model), 32 PCs described 99.5% of the total variance in the data set. One can think of variance as the overall structure of the data. Thus, the 292 dimensions excluded collectively only accounted for 0.5% of the data's structure. This 32D data set was then subject to MLG-LCA, also called linear discriminant analysis (LDA, cf. [9]), in order to quantitatively probe the differences between the accidental patterns generated by each shoe. Table 1 shows the hold-one-out cross validation results for the correct classification of each accidental pattern using MLG-LCA. The overall hold-one-out correct classification rate was 92% (97% apparent correct classification rate). The average hold-one-out correct classification rate was 92 ± 9%. We were surprised at these high correct classification rates, especially considering the fact that evaluation of the data did not include details of the accidentals (such as size and shape) or something like the outsole topography.

The 32D structure of this data is obviously too high in dimensionality to plot. It is none-the-less very instructive to have a physical picture of the data. For this reason projection of the 116 accidental patterns into three-dimensional (3D) PC-space is plotted in Fig. 2 which accounts for 59.7% of the data's variance. While these first three PCs only account for a small portion of the overall

TABLE 1—*Hold-one-out cross-validation table for maximum likelihood Gaussian linear classification of the accidental patterns examined in this study.*

Shoe ID	No. Accidental Patterns Recorded*	No. Misidentified Patterns for Shoe	Incorrectly Predicted Shoe	Individual Shoe "Hold-One-Out" Correct Identification Rates (%)
P1L	12	0		100
P1R	11	0		100
P2L	12	1	P1R	92
P2R	12	0		100
P3L	11	2	P1L, P2L	82
P3R	15	2	P4R, P5R	87
P4L	15	0		100
P4R	13	3	P1L, P5R × 2	77
P5L	0	Omit	Omit	Omit
P5R	15	1	P3R	93

The 324-dimensional accidental patterns were reduced to 32 dimensions (99.5% of total variance) using PCA. This table shows that the average correct identification rate for accidental patterns found on these shoes is estimated to be 92%.

*The accidental patterns used in the classification analysis were those that had at least one accidental mark.

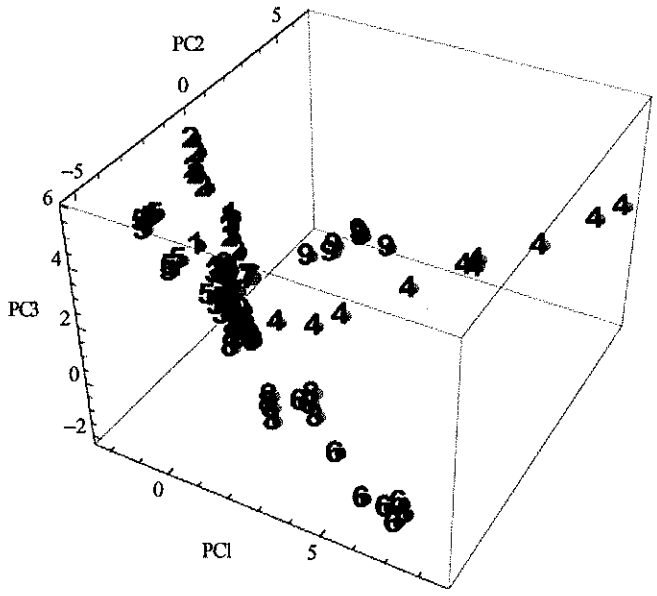


FIG. 2—*All accidental patterns with at least one accidental mark, projected into the space of the first three PCs (59.7% of total variance). The numbers adjacent to each data point label the shoe.*

variance in the data, some clustering of the accidental patterns for particular shoes is evident. This is consistent with the findings in 32D PC-space, i.e., most of the patterns generated by the same shoe are in close proximity. Within the classification theory we employ in this study, MLG-LCA, the more proximate data points are in space, the more likely their identity is the same, i.e., the more likely they are drawn from the same distribution (28,32).

In general, the accidental patterns appear to evolve by starting as a clump of points (i.e., patterns with only one or two accidental marks) and then spreading out linearly in the PC1-PC2 plane (cf. Fig. 2). Interestingly, the accidental patterns of shoes 2 (P1R), 4 (P2R), 5 (P3L), 6 (P3R), 8 (P4R), and 9 (P5R) trace out fairly linear paths in 3D PC-space (Alternative viewpoints of Fig. 2 are available from the authors upon request.) Patterns from shoe 4 (P2R), while somewhat spread out are nonetheless strikingly distinct from the other accidental patterns. These patterns seem to follow a very linear path through 3D PC-space as they change over time. Shoe 9 (P5R) shows accidental patterns that are much closer to each other than those for shoe 4 (P2R) and are also distinct from most of the other patterns. Unfortunately, we cannot infer much about the linear paths traced out by some of the patterns as they evolve over time since these 3D plots account for a relatively low amount of the data's overall variance. The intention of these figures is only to examine if the accidental patterns for the same shoe are relatively close together in 3D PC-space and form distinct clusters.

PCA of Accidental Patterns from Days 14 to 30

The accidental patterns from the first 7 days contained few or no accidental marks. Thus, these patterns are necessarily similar and all tightly clustered (cf. lower left of Fig. 2). When these patterns are removed and the remaining 63 accidental patterns from days 14 to 30 are dimensionally reduced with PCA, the first 28 PCs account for 99.5% of the data's variance. Table 2 shows the hold-one-out cross validation results for correct classification of these accidental patterns using MLG-LCA. The overall hold-one-out

TABLE 2—*Hold-one-out cross-validation table for maximum likelihood Gaussian linear classification of the accidental patterns for days 14–30.*

Shoe ID	No. Accidental Patterns Recorded	No. Misidentified Patterns for Shoe	Incorrectly Predicted Shoe	Individual Shoe "Hold-One-Out" Correct Identification Rates (%)
P1L	7	0		100
P1R	7	2	P2R × 2	71
P2L	7	0		100
P2R	7	0		100
P3L	7	2	P1L × 2	71
P3R	7	1	P4R	86
P4L	7	0		100
P4R	7	0		100
P5L	0	Omit	Omit	Omit
P5R	7	0		100

The 324-dimensional accidental patterns were reduced to 28 dimensions (99.5% of total variance) using PCA. This table shows that the average correct identification rate for accidental patterns in this time period, is estimated to be 92%. This is consistent with the correct identification rate derived from Table 1.

correct classification rate was 92% (100% apparent correct classification rate). The average hold-one-out correct classification rate was 92 ± 13%. These rates are in general quite good although the individual correct classification rates for shoes 2 (P2R) and 5 (P3L) are at 71%. Considering that there are only seven patterns for each shoe, just one misidentification will strongly impact the shoe's average correct classification rate. Given the good overall correct classification rates obtained for this data we would expect the rates for shoes 2 (P2R) and 5 (P3L) would increase if more accidental patterns had been recorded between days 14 and 30.

The first three PCs accounted for 63.6% of the variance for the accidental patterns from days 14 to 30 and their projection into 3D PC-space is shown in Fig. 3. Even at this relatively low variance, accidental patterns for shoes 2 (P1R), 3 (P2L), 5 (P3L), and 9

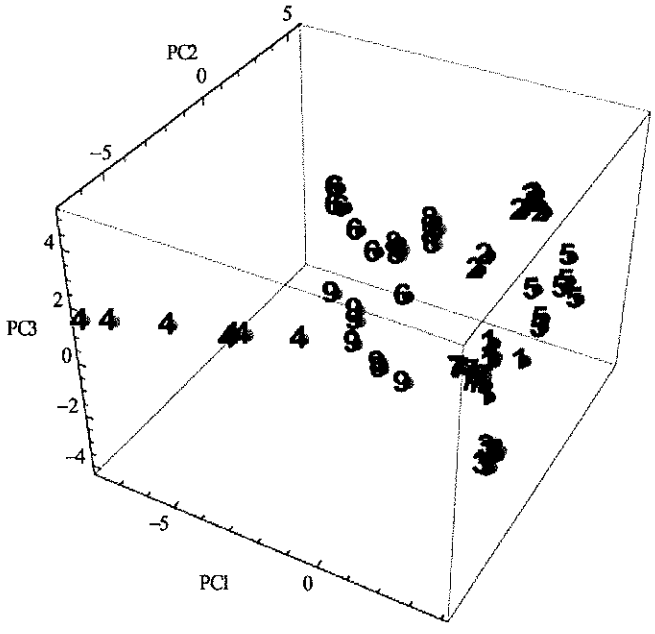


FIG. 3—*Accidental patterns for days 14–30, projected into the space of the first three PCs (63.6% of total variance). The numbers adjacent to each data point label the shoe.*

TABLE 3—*Hold-one-out cross-validation table for maximum likelihood Gaussian linear classification of the accidental patterns for days 20–30.*

Shoe ID	No. Accidental Patterns Recorded	No. Misidentified Patterns for Shoe	Incorrectly Predicted Shoe	Individual Shoe "Hold-One-Out" Correct Identification Rates (%)
P1L	4	0		100
P1R	4	2	P2R × 2	50
P2L	4	0		100
P2R	4	0		100
P3L	4	2	P1L × 2	50
P3R	4	2	P1L, P4R	50
P4L	4	1	P2L	75
P4R	4	0		100
P5L	0	Omit	Omit	Omit
P5R	4	0		100

The 324-dimensional accidental patterns were reduced to 28 dimensions (99.6% of total variance) using PCA. This table shows that the average correct identification rate for accidental patterns in this time period is estimated to be 81%. See text for discussion of this much lower average correct identification rate.

(P5R) are clearly distinct from each other and easy to pick out by eye. Also, when viewing Fig. 3 straight up the PC1 axis (not shown) shoes 6 (P3R) and 8 (P4R) clearly form distinct clusters. The accidental patterns for shoe 7 (P4L) are intermingled, however, with those for shoe 1 (P1L) from any point of view.

PCA of Accidental Patterns from Days 14 to 20

Next, we examine the 36 accidental patterns from days 14 to 20. The first 21 PCs account for 99.6% of the variance in this data set. Table 3 shows the hold-one-out cross validation results for classification of these accidental patterns using MLG-LCA. The overall hold-one-out correct classification rate was 81% (100% apparent correct classification rate). The average hold-one-out correct classification rate was 81 ± 24%. Note that while the overall and average correct classification rates for this data set are low there are only four patterns for each shoe. Each misidentification by MLG-LCA would thus be expected to impact these correct classification rates by a wide margin.

Figure 4 shows a plot of the first three PCs for accidental patterns from days 14 to 20. The plot accounts for 65.7% of this data set's variance. Good clustering of patterns (points) stemming from the same shoe can be seen during this third week of wear. The patterns of shoe 3 (P2L) appear intermingled with those for shoe 1 (P1L). However, if Fig. 3 is viewed straight down the PC1-axis one would see that this is not the case. Similarly, viewing the data straight down the PC3-axis reveals that the patterns for shoes 5 (P3L), 6 (P3R) and 8 (P4R) are all distinct and well separated in space (alternative views of Fig. 3 are available from the authors upon request). Only shoes 1 (P1L) and 7 (P4L) are too close to visually differentiate in 3D PC-space.

PCA of Accidental Patterns from Days 20 to 30

The best classification results of accidental patterns were obtained for the last recorded week of wear, days 20–30 (36 patterns). From a footwear examiners point of view, this is not unexpected as the shoes have accumulated the most wear and therefore have developed the most elaborate patterns of random accidental marks. The first 22 PCs accounted for 99.5% of the data's variance. All hold-one-out cross validation results for correct classification of these accidental patterns using MLG-LCA in 22D PC-space were 100%.

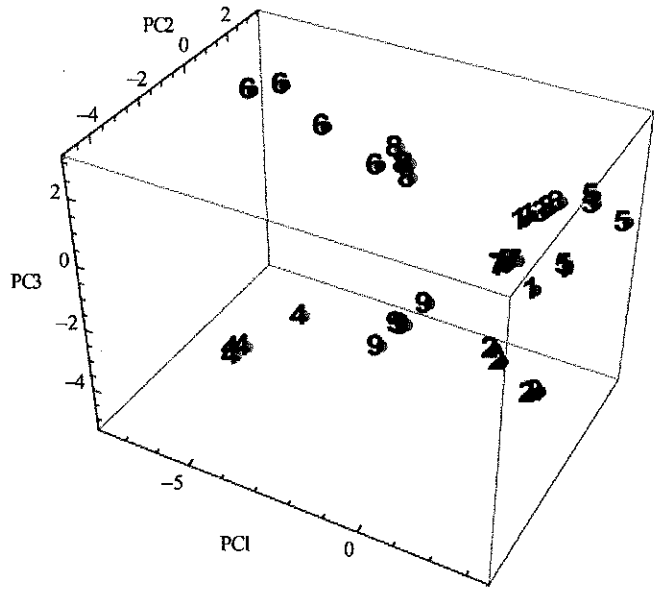


FIG. 4—*Accidental patterns for days 14–20, projected into the space of the first three PCs (65.7%). The numbers adjacent to each data point label the shoe.*

For an approximate, although visual representation of how different the accidental patterns stemming from each shoe are at this point, the data from days 20 to 30 projected into the space of the first three PCs in Fig. 5 (accounts for 66.7% of the data's variance). Even at this relatively low variance value Fig. 5 conveys that the patterns for each shoe form absolutely distinct and well-separated clusters in 3D PC-space. Overall, as time passes it is exquisitely clear that the accidental patterns developed on the outsoles of these

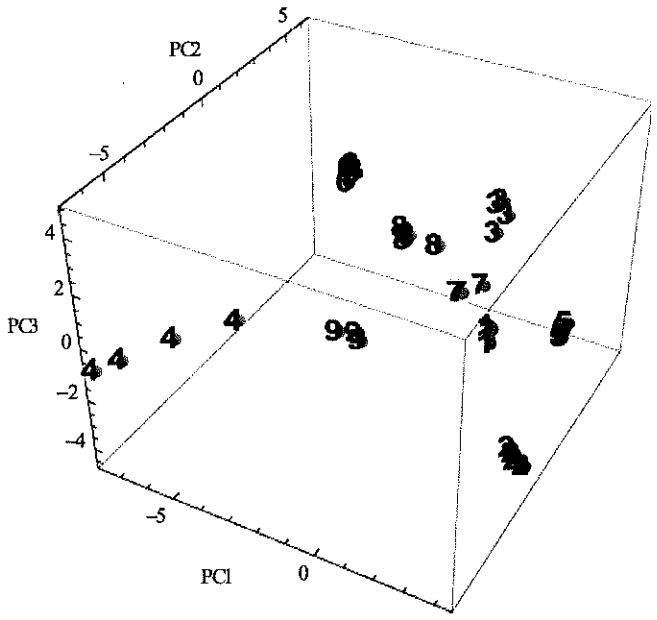


FIG. 5—*Accidental patterns for days 20–30, projected into the space of the first three PCs (66.7% of the total variance). The numbers adjacent to each data point label the shoe.*

shoes became more and more distinguishable, i.e., the inter-cluster spread in data space increases over time.

Conclusion

The footwear accidental pattern comparison technique presented in this study utilized only a tiny amount of the information typically available to the footwear examiner and yet it was usually able to correctly identify which shoe generated a particular pattern. The way the method works is to mathematically compare the distribution of accidental marks (accidental patterns) on the sole of an unknown shoe to multiple accidental patterns generated by shoes of known identity. The statistical comparison method used in this study was maximum likelihood Gaussian linear classification. The identity of the unknown pattern is assigned to the known shoe with statistically the most similar accidental patterns.

The high correct classification rates from our minimally detailed data lend a great deal of credence to the proposition postulated by imprint examiners of the "uniqueness" of accidental patterns. If data are also recorded for the physical characteristics of each accidental, the above results indicate that this method would be even more successful in identifying a shoe from one or more related accidental patterns.

Patterns from the same shoe although at different points in time tended to cluster closer to each other than patterns from different shoes. This was demonstrated (numerically) in high dimensional PC-space using MLG-LCA, and (graphically) in 3D PC-space. The 3D PC-space plots graphically show that generally there is no relationship between the patterns of the left and right shoes from the same pair, as might be expected. If there were such a relationship then one would expect to see tighter clustering between the patterns (i.e., points in the 3D plots) from the same pair of shoes.

Correct classification rates using MLG-LCA and the hold-one-out procedure ranged from 81% to 100% when 99.5% of the data's variance was retained. Two factors affected the correct classification rates, length of time the shoe was worn and the number of accidental patterns included in the analysis. Most notably, the longer the shoe was worn, the more different the patterns became. Although some of this information is well known to the trained footwear examiner, in a court of law if one can use sophisticated yet understandable statistical methods to draw conclusions about evidence and discuss statistical certainties, one can make a profound impression on the courts and support expert footwear examiner testimony that has been arrived at qualitatively.

By using the same manufacturer and model of shoe as well as having the same person as the wearer, many variables that contribute to the "unique" characteristics formed on shoe soles were eliminated or muted. Hence, the ability to still easily distinguish between such shoes with minimally detailed data strongly supports the claims of the great discrimination power of footwear impressions. Logic then dictates that the inclusion of accidental mark details, such as size and shape, will further add to this method's discriminating power.

We have already begun to expand upon this study by increasing the number of participants, frequency of data collection, and lengthening the total time over which accidental patterns are collected. We believe this study will further strengthen and elaborate on our findings here. In the future, we would like to implement automated methods of data collection for the shoes, in particular using high resolution laser scanning to map the surface topography of the outsole. Computer aided design software could then be used to make unbiased measurements and projections of the data for comparison to other shoe soles and shoe sole impressions.

Acknowledgments

We thank Ms. Helen Chan, Mr. Manny Chapparo, Mr. Peter Diaczuk, and Ms. Marta Ekstrom for many valuable discussions and for reading our manuscript.

References

1. Bodziak W. Footwear impression evidence. 2nd edn. Boca Raton, FL: CRC Press, 1995.
2. Cassidy MJ. Footwear identification. 1st edn. Salem, OR: Lightning, 1980.
3. Deskiewicz K, editor. Schallamach pattern on shoe outsole acknowledged by court in footwear identification. Proceedings of the International Symposium on Setting Quality Standards for the Forensic Community; 1999 May 3-7; San Antonio, TX. Washington, DC: Forensic Science Communications (FBI), 1999.
4. Davis RJ, Keeley A. Feathering of footwear. Sci Justice 2000; 40(4):273–6.
5. *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579. (1993).
6. Bohan TL, Heels EJ. The new scientific evidence "standard" and the standards of several states. J Forensic Sci 1995;40(6):1030–44.
7. *United States v. Mahone*, 453 F.3d 68 (1st Cir. 2006).
8. Vapnik VN. Statistical learning theory. 1st edn. New York, NY: Wiley, 1998.
9. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer, 2001.
10. Duda RO, Hart PE, Stork DG. Pattern classification. 2nd edn. New York, NY: Wiley, 2001.
11. Theodoridis S, Koutroumbas K. Pattern recognition. 3rd edn. San Diego, CA: Academic Press, 2006.
12. Everett IW, Lambert JA, Buckleton JS. A Bayesian approach to interpreting footwear marks in forensic casework. Sci Justice 1998;38(4): 241–7.
13. Geradts Z, Keijzer J. The image-database REBEZO for shoeprints with developments on automatic classification of shoe outsole designs. Forensic Sci Int 1996;82:21–31.
14. Geradts Z, Keijzer J, Keereweer I, editors. Automatic comparison of striation marks and automatic classification of shoe marks. Proceedings of SPIE; 1995 July 9; San Diego, CA. Washington, DC: SPIE, 1995.
15. Geradts Z. Content-based information retrieval from forensic image databases [dissertation]. Utrecht: University of Utrecht, 2002.
16. Kirby M. Geometric data analysis: an empirical approach to dimensionality reduction and the study of patterns. 1st edn. New York, NY: Wiley, 2000.
17. Trucco E, Verri A. Introductory techniques for 3D facial recognition. 1st edn. New York, NY: Prentice-Hall, 1998.
18. Grenander U, Miller M. Pattern theory: from representation to inference. 1st edn. New York, NY: Oxford, 2007.
19. Fukunaga K. Statistical pattern recognition. 2nd edn. San Diego, CA: Academic Press, 1990.
20. Turk M, Pentland A. Eigenfaces for recognition. J Cognitive Neurosci 1991;3(1):71–86.
21. Jolliffe IT. Principal component analysis. 2nd edn. New York, NY: Springer, 2004.
22. Burger W, Burge MJ. Digital image processing: an algorithmic introduction using Java. 1st edn. New York, NY: Springer, 2008.
23. Gonzalez RC, Woods RE. Digital image processing. 3rd edn. Saddle River, NJ: Pearson, 2008.
24. Wilshire B, Hurley N. Development of two-dimensional footwear impressions using magnetic flake powders. J Forensic Sci 1996;41(4): 678–80.
25. Abbott JR. Footwear evidence. Springfield, IL: Charles C. Thomas, 1964.
26. Kennedy RB, Pressmann IS, Chen S, Petersen PH, Pressman AE. Statistical analysis of barefoot impressions. J Forensic Sci 2003;48(1):55–63.
27. Kennedy RB, Chen S, Pressmann IS, Yamashita AB, Pressman AE. A large-scale statistical analysis of barefoot impressions. J Forensic Sci 2005;50(5):1071–9.
28. Rencher AC. Methods of multivariate analysis. 2nd edn. Hoboken, NJ: Wiley, 2002.
29. Wolfram Research, Inc. Mathematica [computer program]. Version 5.1. Champaign, IL: Wolfram Research, Inc. 2005.

30. Petraco NDK, Gil M, Pizzola PA, Kubic TA. Statistical discrimination of liquid gasoline samples from casework. J Forensic Sci 2008;53(5): 1092–101.
31. Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. 1st edn. London, UK: Cambridge University Press, 2004.
32. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. 1st edn. Amsterdam, Holland: Academic Press, 1980.
33. Lanyon SM. Jackknifing and bootstrapping: important "new" statistical techniques for ornithologists. Auk 1987;104:144–6.

Additional information and reprint requests:
Nicholas D. K. Petraco, Ph.D.
Department of Science
John Jay College of Criminal Justice
899 10th Avenue
New York
NY 10019
E-mail: npetraco@jjay.cuny.edu

premise seems to have always held. Although the individual characteristics that are present in a shoe print are not necessarily of natural origin, the physical processes by which they are created are certainly random. When these characteristics are considered in combination with other defects or when the internal details or conformation of a single defect is sufficiently complex, the print is considered to be unique.

The defects on the sole of a shoe may be described generally as nicks, scratches, cuts, punctures, tears, embedded air bubbles caused by manufacturing imperfections, and ragged holes. Additionally, there may be foreign materials or particles adhering to the sole or wedged into gaps in the tread pattern elements. Such materials may include pebbles, glass, small sections of twigs, thumbtacks, nails, chewing gum, tar, and adhesive materials (e.g., Shoe Goo, Bostik Shoe Repair Adhesive, McNett Freesole Shoe Repair) designed to repair defects on shoe soles.

The value of a specific defect as an identifier is directly correlated with its dimensional complexity. The more complex the outline of an accidental defect, the less likely it could be duplicated by random processes. The variables associated with such defects may be summarized as follows.

Variables

- *Position*: A defect is characterized by its position on the sole of the shoe. The determination of position may be made relative to the perimeter of a shoe print, relative to particular tread elements or portions of patterns, or relative to other defects.
- *Configuration*: The simplest defect will be referred to as a point. Similar to a point in geometry, a point has no configuration – no discernible shape or elongation. However, if a defect is anything other than a point, it will have some particular shape. It will have a certain length (e.g., a straight-sliced cut in the sole). It may have both length and width (e.g., a 15 mm-long, curving cut in the sole may either form a very shallow curve or a more arced curve). It may have a distinctive, two-dimensional outline (e.g., an irregularly-shaped pebble lodged in a crevice of the tread pattern or a ragged-edged hole worn through the outsole).

- *Orientation*: Excluding point characteristics, a defect of any particular shape will have a specific rotational orientation, differentiating it from another similarly shaped defect that has some different angular orientation. For example, a 25 mm-long, straight-sliced cut in a sole may be parallel with a line from toe to heel, or perpendicular to that line, or rotated to any intermediate angle.

In the course of a shoe print examination, these variables are considered independently for each defect and then in combination with all the other defects. These individual characteristics, along with the class characteristics, enable an examiner to determine the identity or nonidentity of a shoe print when compared with similar characteristics on a suspect shoe.

Hypothetical Shoe and Print

To begin a probabilistic analysis of these types of individual characteristics, a hypothetical shoe and its related shoe print is used with the following assumptions:

- The shoe is a flat-soled, athletic shoe, men's size 8 ½ (US).
- The surface area of a print made by the shoe is 16,000 sq mm (Figure 1).

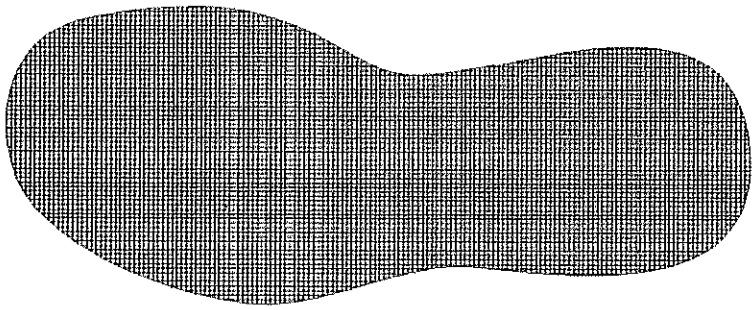


Figure 1

Hypothetical shoe with 16,000 sq mm grid.

- It is equally likely for a defect to appear at any particular position on the shoe as any other. (There may, in actuality, be a greater likelihood for a defect to be present at certain locations than others. For example, defects may more often be found on the heel rather than under the arch of the foot, but for the purposes of this study, it is assumed that no position on the shoe is either more or less likely than any other to have a defect.)

- The position of a defect and the internal details of a defect, if any, may be visualized and measured to a resolution of 1 mm.

- All defects, their conformation, and their orientations are accidental and random (no class characteristics are considered).

- It is acknowledged that all actual defects are, in fact, three-dimensional. However, for the purposes of this study, it is assumed that if a particular dimension does not exceed the minimum resolution (1 mm), it will be treated as if it were two-dimensional. For example, a straight-line cut on the sole of a shoe is 12 mm long and at no point along its length does it exceed 1 mm in width. The real depth and the real width of that cut will both be excluded from consideration.

Standardized Individual Characteristics (Figure 2)

- *Point*: The simplest defect is a point characteristic. It might, for example, be the result of the penetration of the sole by a thorn or a thumbtack that subsequently pulled out and left a minute defect in the sole. A point characteristic has no discernible length or width (or, at least, none that exceed the specified resolution of 1 mm). It has a specific position only.

- *Line*: A line characteristic is straight and has a specific position, a discernible length, and an observable orientation.

- *Curve*: A curve characteristic has a specific position, a discernible length, an observable orientation, and a particular degree of curvature with the apex of the curve located at a specific position on the curve (explanations of these details will follow).
- *Enclosure*: An enclosure characteristic has a specific position, discernible length and width, an observable orientation, and some particular outline shape in two dimensions. It forms an enclosed area around a blank or relatively unmarked space.
- *Three-dimensional*: A three-dimensional characteristic has a specific position, discernible length and width, an observable orientation, some particular outline shape, and what might be thought of as variations in elevation within the interior of the enclosed area.

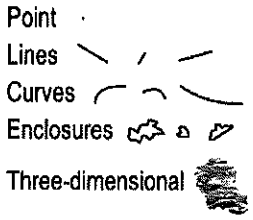


Figure 2

Standardized individual characteristics.

Probabilities of Point Characteristics

A single point characteristic may appear randomly at any one of the 16,000 discrete locations of the shoe print. The formula for simple probability can be used to describe it:

Where P_e = the probability of an event occurring
m = the number of ways of success
n = the total number of possible outcomes

$$P_e = \frac{m}{n}$$

For a single point characteristic (*1 pc*):

$$P_{1pc} = \frac{1}{16{,}000} = .000625$$
$$= 1 \text{ out of } 16{,}000$$

When two (or more) point characteristics occur, there are a finite number of distinct ways the defects could appear on any of the 16,000 positions on the sole. The appropriate formula for a simple combination is:

$$_nC_r = \frac{n!}{(n-r)!\,r!}$$

Where $_nC_r$ = the combination of r items
taken n at a time ["n!" represents "n factorial" which is: $(n \cdot (n-1) \cdot (n-2) \cdot \cdot \cdot 2 \cdot 1)$
For example, $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$]

For two point characteristics the number of distinct combinations is:

$$_{16{,}000}C_2 = \frac{16{,}000!}{(16{,}000-2)!} = \frac{16{,}000 \times 15{,}999}{2} = 127{,}992{,}000$$

# of point characteristics	Probability of random duplication
1	1 out of 16,000
2	1 out of 127,992,000
3	1 out of 6.83E+11
4	1 out of 2.73E+15
5	1 out of 8.73E+18
6	1 out of 2.33E+22
7	1 out of 5.32E+25
8	1 out of 1.06E+29
9	1 out of 1.89E+32
10	1 out of 3.02E+35

Table 1

Random duplication of "point" characteristics.

The odds, then, against looking at another similar theoretical shoe – also with two random point defects – and finding them in the same two positions is 1 out of 127,992,000. (Visualizing numbers with large exponents can range from difficult to impossible.*)

Probabilities of Line Characteristics

A line characteristic is defined as a straight line. Although a line may conceivably start at any one of the 16,000 positions on the shoe sole and may end at any one of the 15,999 remaining positions, practical experience reveals that long lines that continue across a significant portion of the shoe print are encountered far less frequently than shorter lines. The length

* When flying in an aircraft above the surface of the Earth, the immensity of the planet on which we live is well illustrated. One can imagine that the entire sphere of the Earth consists entirely of fine grains of white beach sand – the oceans, crust, mantle, and all the way through the core – nothing but white sand. If a cubic centimeter can contain 8,000 grains of sand, there would be approximately 8,665,655,334,766,030,000,000,000,000,000 grains inside a sphere the same size as the Earth. If a single grain of red sand were randomly inserted within an Earth-sized pile of white ones, the chance of blindly plucking that red grain out would be 1 out of that number of grains – that is, 1 out of 8.666E+30 (8,666 billion billion billion).

of a line, therefore, will be described as short, medium, or long, relative either to other lines within the print or to lines found on shoe prints in general. For example, a short line may be 6 mm or less in length. A medium line may be from 7 mm to 15 mm long, and a long line may be all lines 16 mm or longer in length. These arbitrary divisions – short, medium, or long – are used instead of actual lengths to preserve the conservatism of the approach. Additionally, a line will be described by its rotational orientation. Lines with any of eight orientations may be easily differentiated (Figure 3).

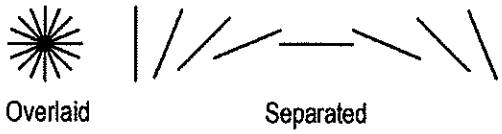


Figure 3

Eight possible orientations for "lines".

These factors for lines, multiplied by the factor for position, yield the following:

3 (length) x 8 (orientation) x 16,000 (position) = 384,000

The probability of random duplication of a single line characteristic is 1 out of 384,000.

# of line characteristics	Probability of random duplication
1	1 out of 384,000
2	1 out of 7.37E+10
3	1 out of 9.44E+15
4	1 out of 9.06E+20
5	1 out of 6.96E+25
6	1 out of 4.45E+30
7	1 out of 2.44E+35
8	1 out of 1.17E+40
9	1 out of 5.00E+44
10	1 out of 1.92E+49

Table 2

Random duplication of line characteristics.

Probabilities of Curve Characteristics

A curve characteristic has the same attributes as a line with the addition of an observable curvature. The analysis will be limited to curves that are relatively shallow. This restriction permits the consideration of curves that approach or equal arcs of circles but excludes all curves that are more curved than that – elongated parabolic curves, for example (Figure 4). Based on experience in the examination of actual shoe prints, elongated curves are encountered far less frequently than more shallow ones.

Because curves share the fundamental factors of lines, the determination of the total possible number of curves begins there (384,000 possibilities). For each one of those curves, the direction of curvature may be either arbitrarily positive (upward or to the right) or arbitrarily negative (downward or to the left).

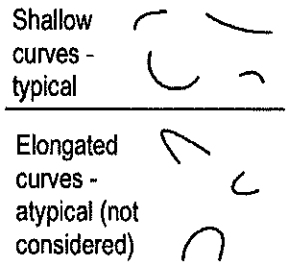


Figure 4

Shallow vs elongated curves.

Additional attributes that differentiate curves from one another were assigned (Figure 5). A curve will exhibit a certain degree of curvature (barely curved, very shallow, shallow, almost an arc, arced). When viewing curves from an arbitrary but consistent viewpoint (for example, curving upward with the end points below the peak), the location of the point of highest curvature, the apex of the curve, may vary (far to the left, left of center, centered, right of center, far to the right).

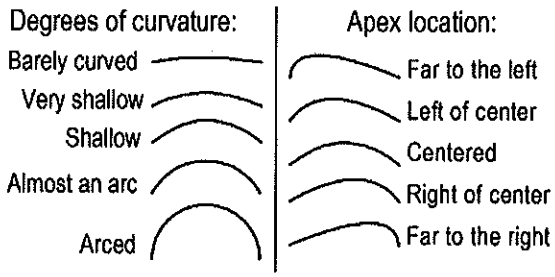


Figure 5

Degrees of curvature and apex location.

The rationale behind these attributes is that an examiner could easily differentiate two curves in which either the degree of curvature was different with the apex location the same or in which the apex location was different with the same degree of curvature. They would be distinguishable in both cases as curves with different shapes.

The combined factors for a curve characteristic are:

16,000 (position) x 3 (length) x 8 (orientation)
x 2 (direction of curvature) x 5 (degree of curvature)
x 5 (apex location) = 19,200,000

If the center point of a particular curve lies at the edge of the theoretical shoe print, that curve can only curve in one direction toward the center of the print, because curving away from the center would place the points that constitute the curve outside the perimeter of the print. This eliminates a number of possible curves. Also, if a curve is very short, there may not be sufficient length for the curve to exhibit a discernible apex location. These limitations to the number of total possible curves have been offset

by limiting the attributes to only eight possible orientations, to only five degrees of curvature, and to only five apex location classes. Certainly each of these distinguishing features could be differentiated to greater degrees than the specified number of varieties, more than compensating for the reduction in number due to the exclusion of curves that would be out of bounds or too short. Additionally, different geometric types of curves that could be distinguished and classified (circular, parabolic, hyperbolic, elliptical, spiral) have, likewise, not been considered. Excluding these subclasses of curves as factors further increases the conservatism.

Finally then, if a single curve characteristic appears on a shoe print, the probability of another print bearing a random, accidental, curve characteristic of the same length and orientation, with the same degree and direction of curvature, with the same apex location, and in the same position on the print is 1 out of 19,200,000.

# of curve characteristics	Probability of random duplication
1	1 out of 19,200,000
2	1 out of 1.84E+14
3	1 out of 1.18E+21
4	1 out of 5.66E+27
5	1 out of 2.17E+34
6	1 out of 6.96E+40
7	1 out of 1.91E+47
8	1 out of 4.58E+53
9	1 out of 9.77E+59
10	1 out of 1.88E+66

Table 3

Random duplication of curve characteristics.

Probabilities of Enclosure Characteristics

Two-dimensional enclosure characteristics (referred to in geometry as curvilinear shapes) range from relatively simple to very complex in outline. Figure 6 shows some examples.

Figure 6

Enclosure characteristics.

Initial thoughts about characterizing enclosures were that they consisted of a specific composite of single points along the perimeter of the enclosure, and that the probability of those specific points, taken in combination, would define the enclosure. This would have made duplication by random processes of even relatively simple enclosures highly unlikely. But assumptions about how enclosures are presumably created led to what is believed to be a more realistic (and mathematically conservative) conclusion as to how they should be modeled.

Three primary possibilities exist for creating an enclosure-type defect on a shoe sole. First, a sole may be defaced by some small object that is the same general shape as the resultant defect (though reversed or negative). Second, a section of the sole may be physically torn away from the rest of the sole. And third, severe wear of the sole may penetrate the exterior outsole layer, leaving an enclosure-type window through that exterior layer to the midsole or underlayment. It seems an approach that takes these mechanisms of origin into account is in order. Therefore, each enclosure will be treated as an entity, rather than as an agglomerate of individual, random points.

An enclosure may be one of two types. A geometric enclosure would be the outline of some simple geometric shape – a circle, an oval, a triangle, a square, a rectangle, and so forth. An irregular enclosure would bear an outline that would be described as jagged, asymmetric, or random. The four enclosures portrayed

in Figure 6 would be of this type. Although geometric enclosures could certainly originate from random, accidental damage to a shoe, the likelihood that they were created by damage resulting from contact with some geometric-shaped object is greater. This reduces their value as identifiers when compared with irregular enclosures. For example, the shoes of construction workers might be marked by similarly shaped, two-dimensional rectangles as a result of the workers having stepped on the edges of the heads of nails scattered around the job site. Or, the shoes of workers in a machine shop might be marked by similarly shaped crescents from walking on metal shavings from lathes. In either case, the locations and the orientations of the defects on the shoes would be random, but the replication of the geometric shapes themselves would not be.

A geometric enclosure in the shape of a circle would exhibit no distinguishable orientation. Circular enclosures will be characterized by their size and location only:

3 (size) x 16,000 (position) = 48,000

# of circular enclosure characteristics	Probability of random duplication
1	1 out of 48,000
2	1 out of 1.15E+09
3	1 out of 1.84E+13
4	1 out of 2.21E+17
5	1 out of 2.12E+21
6	1 out of 1.70E+25
7	1 out of 1.16E+29
8	1 out of 6.98E+32
9	1 out of 3.72E+36
10	1 out of 1.79E+40

Table 4

Random duplication of circular enclosure characteristics.

Noncircular geometric enclosures will be characterized by their relative size, their orientation, and their location. The combined factors for a noncircular geometric enclosure characteristic are:

3 (size) x 8 (orientation) x 16,000 (position) = 384,000

# of non-circular geometric enclosure	Probability of random duplication
1	1 out of 384,000
2	1 out of 7.37E+10
3	1 out of 9.44E+15
4	1 out of 9.06E+20
5	1 out of 6.96E+25
6	1 out of 4.45E+30
7	1 out of 2.44E+35
8	1 out of 1.17E+40
9	1 out of 5.00E+44
10	1 out of 1.92E+49

Table 5

Random duplication of geometric enclosure characteristics.

Reasonably, an irregular enclosure with a more complex outline will have greater value as an identifier than one with a simple outline. They may be characterized by the number of directional deviations along their perimeter. For example, the irregular enclosure in Figure 7 exhibits ten such changes of direction. As one traces the perimeter of the enclosure, beginning between 10 and 1, there is a very sharp deviation to the right at 1, there is a sharp deviation to the left at 2, a short and somewhat rounded deviation to the right at 3, a fairly sharp deviation to the right at 4, a very gradual deviation to the left at 5, and so on. To assign a value to the complexity of such enclosures, the number of possible direction changes (2, either right or left) will be raised to the power of the number of directional deviations (10 for the enclosure in Figure 7). A similar methodology has been employed to characterize two-dimensional fractures [1]. Application of this mathematical model to irregular enclosures disregards the additional variables of the angularity of each deviation (e.g., sharp or gentle) and the dissimilar distances along the contour between adjacent deviations. Two irregular enclosures having the same number of directional deviations but with differing angularities and distances would certainly appear different when compared. Figure 8 exhibits directional deviations that correspond with those in Figure 7 with regard to total number and direction but not in angularity or the distances along the contour between adjacent deviations. Their treatment as equivalent by the quantifying methodology adds significantly

to the conservatism of the approach. This treatment, 2 raised to the power of the number of directional deviations, will be referred to as the complexity factor.

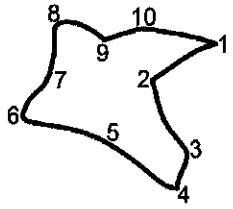


Figure 7

Ten directional deviations of an irregular enclosure.

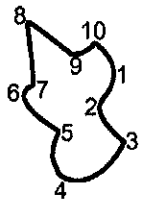


Figure 8

An irregular enclosure with ten similar directional deviations but different angularities and lengths.

Irregular enclosures, then, will be characterized by their relative size, their orientation, their location, and their complexity. The combined factors for an irregular enclosure characteristic are:

3 (size) x 8 (orientation) x 16,000 (position) x (complexity factor)

Because irregular enclosures are less common than the previously discussed types of characteristics, Table 6 lists the probability of random duplication of single irregular enclosure characteristics by complexity factor.

# of directional deviations in an irregular enclosure (complexity factor)	Probability of random duplication of that single irregular enclosure
3	1 out of 3.07E+06
4	1 out of 6.14E+06
5	1 out of 1.23E+07
6	1 out of 2.46E+07
7	1 out of 4.92E+07
8	1 out of 9.83E+07
9	1 out of 1.97E+08
10	1 out of 3.93E+08
11	1 out of 7.86E+08
12	1 out of 1.57E+09
13	1 out of 3.15E+09
14	1 out of 6.29E+09
15	1 out of 1.26E+10
16	1 out of 2.52E+10
17	1 out of 5.03E+10
18	1 out of 1.01E+11
19	1 out of 2.01E+11
20	1 out of 4.03E+11

Table 6

Random duplication of single irregular enclosure characteristics.

Probabilities of Three-dimensional Characteristics

When a portion of a shoe sole, such as the lug of a work boot, is physically broken off, the exposed surface may exhibit random variations in height. Naturally, such a surface requires an impressionable medium – a two-dimensional shoe print will not record these variations. The nature and appearance of the fractured surface is dependent upon the material of which the sole is made, and random, irregularly fractured surfaces may not occur with some shoes. In ideal cases, the exposed surface will resemble the interior surfaces of brittle metals that have been fractured. Fractals have been proposed by Thornton [2] to model these surfaces. His attempts to determine their complexity and "degree of uniqueness" were based on the processing time required by a computer to calculate the fractals.

All of the comments and calculations that follow are based on the assumption that the surfaces are random in nature both in the method of generation and in their appearance.

The more complex the fractured surface, the less likely it is to be duplicated by random production. The modeled surface in Figure 9 depicts a 50 x 50 grid with random surface elevations.

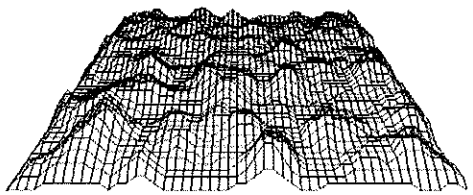


Figure 9

50 x 50 grid with random surface elevations.

An understanding of the method used to determine the probability of occurrence of such a surface is aided by significant simplification. Therefore, the first surface to be examined will be a 3 x 3 grid with elevation variations of either 0 or +1 unit only (Figure 10). Each of the labeled intersections represents a point where the elevation could be either 0 or +1. In the illustrated case, the only +1 point is that labeled 9. The other eight points are at 0 elevation.

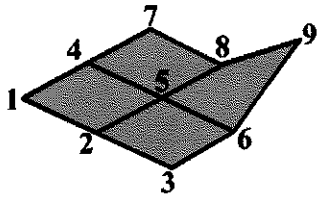


Figure 10

3 x 3 grid.

The formula for complete variations is used to determine the number of possibilities in such a case. The number of possible height variations is raised to the power of the number of positions on the grid. In the simple case above, there are 2^9, or 512, different variations of a model fracture that size that has only two possible variations in height.

Grid size is equivalent to a measurement of the area of the defect. If details within the fracture on the shoe print are resolvable to 1 mm, then the area of the defect would be measured or estimated in square millimeters. That measurement of area becomes the exponent of the number of variations, and the depth, measured using the same resolution, becomes the base number.

Again, to maintain conservatism, the height variation will be limited to only two possibilities.

Grid size	# of variations	Probability of random duplication
2 x 2	2^4	1 out of 16
3 x 3	2^9	1 out of 512
4 x 4	2^{16}	1 out of 65,536
5 x 5	2^{25}	1 out of 33,554,432
10 x 10	2^{100}	1 out of 1.27E+30
15 x 15	2^{225}	1 out of 5.39E+67
20 x 20	2^{400}	1 out of 2.58E+120

Table 7

Random duplication of three-dimensional characteristics.

Figure 11 is an example of such a 10 x 10 grid listed in Table 7 with randomly generated heights of +1 or 0 for each of the 100 points.

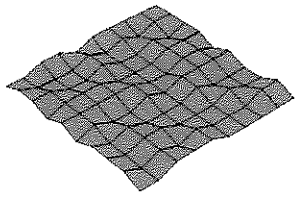


Figure 11

10 x 10 grid with random heights of +1 or 0.

Additional possible height variations, such as those previously depicted in Figure 9, increase the magnitudes of the probabilities tremendously.

Compound Characteristics

It is not uncommon to find what will be termed compound characteristics in a shoe print. These consist of two or more joined (or apparently joined) defects. Probably the most common would be a curving line that consists of multiple curve characteristics. These might be referred to as meandering curves (Figure 12).

Figure 12

Meandering curves.

Other compound characteristics might consist of an enclosure with an attached line or a curve attached to a three-dimensional characteristic. Though such defects are significantly more distinctive than regular characteristics, any quantification of compound characteristics should be applied to the separate, distinct portions of the defects individually to maintain a conservative methodology.

Combinations of Characteristics

Most shoes will exhibit more than one individual characteristic, and commonly some combination of the above types of characteristics will be present. For example, a shoe sole may have two point characteristics, one line characteristic, two curve characteristics, and one irregular enclosure characteristic (with, say, a complexity factor of 6) all distributed randomly on its surface.

A determination of the combined probability of these separate characteristics on the hypothetical shoe requires that the individual probabilities of each of the characteristics be multiplied together. For the example mentioned above, the

probability of random duplication of a single point characteristic is 1 out of 16,000. The second point characteristic could be anywhere on the shoe surface except at the same position as the first point. Therefore, the probability of random duplication of that second point characteristic is 1 out of 15,999. Similarly, neither the line characteristic nor the curve characteristics could include either of the two point characteristic locations (or they would obscure that point completely). A slight modification of the line and curve probabilities must be made to reflect this restriction; they are recalculated as if the surface area were reduced by 1 sq mm for each preceding characteristic. For the hypothetical print that displays two points, one line, two curves, and irregular enclosure of complexity 6, then, the formula for this particular combination of characteristics is:

$$P_{Comb.\ of\ Chars.} = \overbrace{\frac{1}{16{,}000}}^{1st\ point} \times \overbrace{\frac{1}{15{,}999}}^{2nd\ point} \times \overbrace{\frac{1}{383{,}998}}^{line} \times \overbrace{\frac{1}{19{,}199{,}997}}^{1st\ curve} \times \overbrace{\frac{1}{19{,}199{,}996}}^{2nd\ curve} \times \overbrace{\frac{1}{25{,}576{,}000}}^{enclosure}$$

$$= 1.12E\text{-}36$$

$$= 1 \text{ out of } 8.91E{+}35$$

This method of multiplying the individual probabilities together is appropriate for any particular combination of characteristics.

Conclusions

This study is not necessarily meant to reflect the probabilities of finding these types of characteristics in actual shoe prints. By establishing the criteria in what is felt to be a conservative and quasi-realistic manner, this analysis does yield information about the sometimes-incomprehensible magnitude of the "uniqueness" of these types of characteristics when they occur in multiples or combinations.

Any decrease in the effective resolution of the details of a shoe print (e.g., if the print was made in coarse soil) will significantly affect any estimated or theoretical probabilities. A simple example will demonstrate. If the resolution is decreased such that features smaller than 2 mm are blurred or undetectable, the original 16,000 sq mm area is reduced to 4,000 resolvable positions. The number of possible resolvable lines, then, is reduced from 384,000 to 96,000, which is a reduction factor of 4.

If the area of a print is less than 16,000 sq mm (for example, when the shoe is a smaller size), the estimated or theoretical probabilities will also be decreased. However, if only a partial shoe print (of the hypothetical average-sized shoe used herein) is considered, the estimated or theoretical probabilities would not be affected.

Vehicle tire prints were not directly addressed by this study, but the same methodology could be applied. The surface area of an average-sized tire on a compact car is about 280,000 sq mm. Assuming the same 1-mm resolution, the theoretical probability of random duplication of two point characteristics would be 1 out of 39,199,860,000.

The numerical variables associated with the class characteristics of shoes were not included in probability calculations. Consideration of these factors would certainly further individualize a particular shoe or print, in some cases, drastically. There are tens of thousands of different shoe sole patterns. Not only is each pattern available in many different sizes, but also soles may exhibit varying areas and degrees of wear, mold defects, injection-related voids, foxing strips, and heel labels.

Before an examiner could even consider applying probability estimates to actual casework, validation studies would have to be performed to verify that true-life occurrences are accurately modeled. Two scenarios come to mind. A number of new, unmarked, identical pairs of shoes could be obtained. Shoes with flat and relatively unmarked outsoles would be preferred. They could either all be worn by the same individual or by several different individuals to walk across miscellaneous materials that would create marks on them. The resultant marks could then be examined and quantified. A second study could be

structured as an examination of existing defects on old shoes that are acquired from different sources and that had been worn by different people. Any validation study should at least address these questions:

1. Is the position of a characteristic on a shoe (and shoe print) resolvable to 1 mm?

2. Are points and other described characteristics actually found on shoes (and shoe prints)?

3. Is it possible to differentiate the rotational orientation of characteristics to at least eight different angles?

4. Is there greater likelihood of finding characteristics at some locations on shoes as opposed to others (for example, are characteristics more commonly found on heels rather than on instep portions)?

5. Are long line characteristics (those that traverse a significant portion of the length or width of the shoe) less often encountered than shorter line characteristics?

6. Are relatively shallow curves (arcs of circles or less) more or less common than elongated curves?

7. Can the attributes of curves be differentiated by at least five degrees of curvature and at least five apex locations?

8. Are the five listed degrees of curvature and the five apex locations equally likely to occur?

Even infinitesimally small probabilities, in and of themselves, would never directly allow examiners to "positively identify" a print as having been created by a particular shoe. However, such statistics as these provide the examiner (and possibly a judge or the members of a jury) with some perspective regarding the incredible uniqueness of a shoe with even a small number of individual characteristics.

For further information, please contact:

Rocky S. Stone
741 South Highway 217
Tijeras, NM 87059
rockyabq@aol.com

References

1. Stone, R. S. A Probabilistic Model of Fractures in Brittle Metals. *Assoc. of Firearm and Tool Mark Exam. J.* **2004**, *36* (4), 297.
2. Thornton, J. I. Fractal Surfaces as Models of Physical Matches. *J. For. Sci.* **1986**, *31* (4), 1435.

Technical Note

The Mount Bierstadt Study: An Experiment in Unique Damage Formation in Footwear

T. W. Adair[1]
J. Lemay[2]
A. McDonald[3]
R. Shaw[3]
R. Tewes[4]

Abstract: Randomly formed damage on footwear outsoles has appropriately been used to compare crime scene impressions to the known shoes of suspects, witnesses, and victims. In this study, the authors wore new, identical boots (two pairs) during a seven-mile hike. The authors attempted to control the major variables except the manner in which the outsole of the boot made contact with the ground. The results of this experiment support the use of these marks for the individualization of footwear and confirm their random formation through the use of the shoe by the wearer.

Introduction

The use of random damage characteristics has been reliably used in the comparison of known outsoles to questioned impressions found at crime scenes [1, 2]. These damage characteristics are formed through the use of the shoes while they are worn. The presence of these characteristics in both the known shoe and crime scene impression may allow the footwear examiner to individualize one shoe as having made an impression.

Received January 30, 2006; accepted April 26, 2006

1 Westminster Police Department, Westminster, CO
2 Weld County Sheriff's Office, Greeley, CO
3 Arapahoe County Sheriff's Office, Centennial, CO
4 Fort Collins Police Department, Fort Collins, CO

The authors hypothesized that these characteristics could be created during a single common activity, such as a hiking trip. It was further hypothesized that the characteristics would exist in sufficient numbers to individualize the shoes, even among participants wearing the same shoes and following the same general walking path under the same environmental conditions. By exposing the same manufactured shoes to the same environmental, topographical, and duration of use conditions, the authors hoped to test the premise that accidental characteristics would be created, allowing for the individualization of the shoes. To that end, this study was developed to answer the following questions:

1. Will random characteristics be created in sufficient numbers to allow for individualization by low- to medium-impact walking over relatively short distances?

2. Will these random characteristics share any common location or orientation with other characteristics found on outsoles exposed to the same testing conditions but worn by different persons?

3. Would two pairs of shoes worn by the same individual under the same physical conditions exhibit accidental characteristics allowing for individualization?

Materials and Methods

The Altitude II hiking boot from the Hi Tec Corporation was selected as the test shoe in this study (Figure 1). The Altitude II has a seam-sealed waterproof nubuc leather upper with a carbon rubber outsole. It also features a lightweight compression-molded midsole and steel shank. The finished weight of the shoe is about 21oz. Twelve pairs of boots were acquired from the company. The men's sizes were all 10.5 (US), and the women's sizes were 8, 9, and 10 (US). Each of the six participants (three men and three women) had two pairs of boots for the study. One pair was designated to be worn during the ascent and the other during the descent. The boots were not damaged prior to the study. Some boots were briefly worn on carpeting (indoors) prior to the study to break them in but were inspected prior to the hike to ensure that no damage had occurred. Photographs and black fingerprint powder transparency lifts were made of each shoe prior to and

Figure 1

Outsole of the Altitude II hiking boot from the Hi Tec Corporation.

immediately after the hike. One female participant was unable to make the initial hike and completed the hike with one of the authors a few weeks later. The environmental conditions were similar to those experienced by the previous group.

Mount Bierstadt in Colorado has an elevation of 14,065 feet and a trailhead beginning at 11,691 feet. The total hiking time was approximately 3.5 hours and the total distance traveled (round trip) was approximately 7 miles with a 2400-foot gain in elevation from trailhead to summit. Mount Bierstadt is ranked as an "easy" hike and was selected by the authors because it represented a popular destination and relatively low-impact hiking conditions. On July 31, 2005, the authors began their climb of Mount Bierstadt on Guanella Pass near Georgetown, Colorado. The trail is composed primarily of compacted groomed soil with randomly occurring larger rocks embedded in the surface mix. The last part (approximately 200-250 yards) of the trail is comprised of a large boulder field that requires slow and deliberate foot placement. Although the hike was expected to be fairly easy and completed in one day, unpredictable and rapidly changing weather conditions, commonly found at these high elevations, were encountered. The authors experienced a hailstorm lasting nearly the entire descent. Although the trail conditions became wetter during the descent, the compactness of the trail did not seem to differ significantly from the ascent.

Discussion

Shortly after completing the hike, the authors rephotographed all the boots and made transparent black powdered lifts. The boots were not worn between the completion of the hike and the documentation of the outsole conditions. The authors then examined the outsoles for the presence of the accidental characteristics. To assist in the examination of these boots, the individual outsole elements were given a numerical "address" from one to thirty-eight (Figure 2). Examiners noted the presence of each accidental mark and its address (location) on the outsole. The authors counted only the accidental marks that comprised sufficient size and shape to suggest that they would be represented in a crime scene impression under favorable conditions and could be used to individualize the boot. (This study did not seek to determine whether each and every accidental mark would be reproduced in a latent or crime scene impression, because

Figure 2

Outsole of the Altitude II hiking boot with "address" identifiers 1 to 38.

such findings would be predicated on a number of changing factors such as the transfer medium and the surface receiving the impression.) The men's and women's boots differed slightly in the number of outsole elements containing at least one accidental mark. On average, the men's boots contained accidental marks on 44% of the elements, whereas the women's boots contained marks on 33% of the outsole elements. Each male outweighed each female by at least 50 pounds, which may contribute to heavier footsteps, resulting in additional damage. Within the gender groups there was variation as well. One male had as few as 13% of the elements (representing five accidental marks on one boot) with accidental marks, whereas another male had one boot with 62% of the elements containing marks. (The first male subject is a long-time hiker and stated that he subconsciously avoids larger rocks and drop-offs on trails. This may help to explain the lower number on average for this male subject when compared to the other two.)

The outsoles were then compared to each other to determine whether they contained enough detail to be individualized. Each outsole did contain a sufficient number of accidental marks to allow for individualization and each outsole could easily be differentiated from the other outsoles in the study. Additionally, the transparent lifts were reversed and compared as well with the same results. This resulted in a comparison group of 24 outsoles (12 normal, 12 reversed) totaling 576 comparisons. Although significant differences in the physical size of outsoles would normally be sufficient to eliminate a questioned impression, the authors also verified the lack of corresponding damage in the same "address locations" on the outsole, regardless of their physical size differences. In other words, accidental marks found on address location #32 on a women's size 8 outsole were compared to any accidental marks found at the same address location on a men's size 10.5 outsole. No corresponding marks were found at any locations between the subject outsoles (normal or reversed) in this study.

Conclusion

The results of this study demonstrate the widely accepted proposition that the accidental damage found on footwear outsoles is randomly produced. This study attempted to eliminate as many variables contributing to the formation of these

accidental marks as possible. By using the same style of boots (in the same new condition), the same walking path, the same environmental conditions, and the same duration of use, the authors were able to eliminate all major contributing factors to the formation of these marks, aside from the subject's walking style and the random manner in which the outsole made contact with the micro-topography of the walking trail. In addition, the results of this study indicate that these accidental marks may be created by a single walking event, representing one of many changes occurring in the evolution of the damage and wear represented on the outsole. Similar studies should be conducted and reported to further test these findings under varied conditions.

Acknowledgments

The authors wish to thank Cheryl Rebsamen and the Hi Tec and Magnum Boots divisions for their donation of the boots for this study. Your attitude and enthusiasm was a great benefit to all.

For further information, please contact:

Thomas W. Adair
Senior Criminalist
Westminster Police Department
9110 Yates Street
Westminster, CO 80031
303-430-2400 x 4269
tadair@ci.westminster.co.us

References

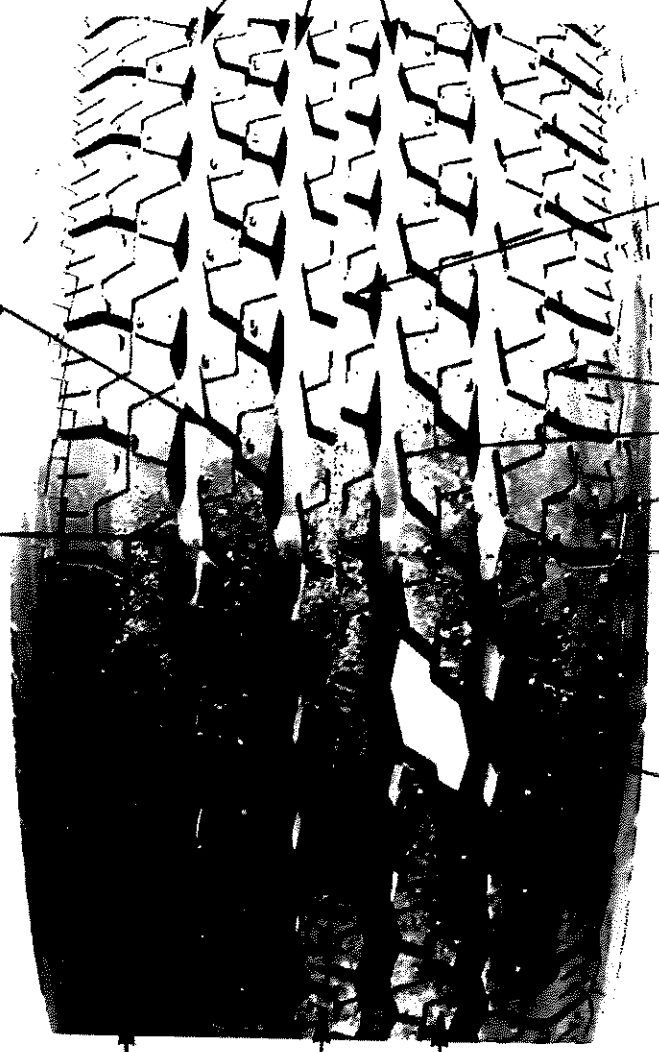
1. Bodziak, W. J. *Footwear Impression Evidence*; CRC Press: Boca Raton FL, 1995.
2. Cassidy, M. J. *Footwear Identification*; Lightning Powder Company: Salem, OR, 1995.

ROYAL CANADIAN MOUNTED POLICE

GAZETTE

Canada

RADIAL TIRE



FIVE RIB TREAD DESIGN

THE SCIENCE OF TIRE IMPRESSION IDENTIFICATION

by Cpl. L.A. Nause, Carlyle Identification Section,
Regina RCMP Sub-Division, Carlyle, Saskatchewan

INTRODUCTION

Today's modern pneumatic tire is the result of almost 100 years of applied science, engineering and research. The tire is composed of many different components each meeting specific requirements, yet all working together to achieve the basic functions of a pneumatic tire.

The tire tread is one such component and it has been the object of much research and development since the very early days of the pneumatic tire.

1. *Tire Technology* — F.J. Kovac, The Goodyear Tire & Rubber Co., page 4.

Tread — The tread is the abrasion resistant component of the tire and forms a protective covering for the carcass. The tread has to be designed for traction, silent running and low heat buildup. The tread is normally composed of a blend of oil-extended SBR and polybutadiene elastomers which have been compounded by adding carbon black, oils, curative ingredients and other chemicals and fillers. A compounded elastomer is popularly called "rubber". The composition of the rubber, the cross section shape of the tread, the number of ribs and grooves and tread design determine the wearing quality, tractions and heat buildup of the tread.[1]

In order to better understand some terminology which will be used later in this article, refer to Figure 1, which points out various components of a modern tread design.

It is the design of the tread which leaves its telltale impression behind at the scene of a crime and is, therefore, the object of much interest among those involved in forensic tire impression identification. In *Tire Tracks and Tread Marks* by Given, Nehrich and Shields the authors state,

Cpl. Nause was born and grew up in Nova Scotia where he joined the RCMP at Halifax in 1971, completing recruit training in Regina, Sask. He was posted to Toronto, Ontario, for one year and then transferred to Niagara Falls, Ontario, where he served from 1972 to 1975.

In the fall of 1975, he was accepted into the Identification Section and posted to Dauphin, Manitoba, in 1976. In 1983 he was transferred to the position of N.C.O. In Charge, Carlyle Identification Section, Carlyle, Saskatchewan.

The author has had other articles relating to the identification field published in the Gazette. In particular an article titled "Tire Impressions as Evidence" was printed in the RCMP Gazette, *Vol. 44, No. 12, 1982. That article was later republished in the Belgium police magazine* Revue de la Gendarmerie *No. 102, 1985 and No. 103, 1986, and in* Identification Canada, *Volume 8, Issue 3, July 1985.*

The article has also been used as hand out material in Canada and the United States on courses dealing with forensic tire impression identification.

Cpl. Nause has been lecturing at the Canadian Police College, Ottawa, Ontario, as a resource person in the field of tire forensics since 1982.

A motor vehicle is used in 75 percent of all major crimes reported today.[2]

It is for this reason that tire impressions can play such an important role in saving investigative manhours as well as providing valuable evidence in court proceedings.

The scientific approach that engineers have used in designing today's modern treads can be used to the advantage of the forensic expert in dealing with crime scene impressions, as you will see later in this article.

CHANGES TO TREAD DESIGNS

In 1845, an Englishman named Robert William Thompson invented a new type of tire which was made of a rubber coated canvas tube covered by leather.

This was the world's first pneumatic (nu-matick) tire which comes from the Greek word, meaning "wind or air".[3]

Not until 1888, however, when the pneumatic tire was reinvented did it begin to achieve wide-spread use. These early tires were completely bald and a motorist was fortunate to travel one hundred miles without a flat. Poor road conditions were a major concern which led to the introduction of the first tread designs to provide some much-needed traction. The year was 1907.

The first traction design was the Firestone tire design using the words "Firestone Non-Skid" as the design, credit is given to the company's founder, Harvey Firestone.[4]

The other manufacturers soon introduced their own tread designs and the attempt to develop new and improved designs has been going on ever since.

Goodyear's first tread design consisted of well-defined, diamond-shaped elements. It was a successful design which they continued to use on passenger tires for many years.

As the condition of the early roads improved and major routes began to be paved these early button-type tread designs used for traction gave way to the continuous rib designs of the 1930's. These designs were more suited to smoother roads and higher speeds. The tread consisted of more-or-less circumferential rows of tread rubber separated by grooves. In order to improve traction on slippery surfaces, the use of sipes, or kerfs as they are sometimes called, were introduced. The main tire construction used during this period in North America was bias ply and it continued to be the most common construction up until the 1960's.

In the mid 1960's the introduction of the belted bias and, in particular, the radial ply construction had a definite effect on tread designs. The radial tire was actually invented in 1913 by Messrs. Gray and Sloper who obtained a British patent for this construction. Although the radial tire was commercially produced as far back as the 1930's it did not gain popular use in North America until the mid 1960's. Since that time the radial tire has continued to capture an ever increasing percentage of the tire market until now it is by far the largest seller of the three basic tire constructions. It appears consumers are prepared to pay the higher cost for radial tires since they provide such superior performance and greater mileage.

Bias ply construction tires did not have continuous rib designs by chance. The early button-type traction designs would not have been able to withstand the higher speeds that came with the changing times. The aggressive traction designs would have destroyed the tread because of the excessive amount of tread squirm in the bias ply construction.

The radial tire construction, however, with its flexible sidewalls and rigid belts greatly improved tread stablization. Less tread squirm has allowed engineers to develop more open and agressive tread designs that can take today's high speeds. In a way, the popularity of the radial tire has seen a return to the early button-type tread designs, although today's designs are much more sophisticated. To appreciate this, one has only to walk into one of the many tire showrooms and examine some tread designs. The radial tire with its well-defined, tread-block shapes and intricate siping treatment is the trend of the future.

The development and increasing popularity of the all-season tread design is yet a continuation of this trend. Goodyear has achieved a great deal of success with its curvilenal tread design on the Vector and we can expect to see similar designs on other tires in the next few years.

NOISE TREATMENT — VARIABLE PITCH

Every person working on forensic tire impression identification should be aware of tread design *noise treatment* and *variable pitch* and how it applies to the comparison process.

As was mentioned already, with improved roads and automobiles came increased highway speeds. The smoother road surfaces and better engineered cars also reduced previous noise levels. Now the sound produced by early constant pitch tread designs became unacceptable.

As a tire rolls over the road surface the tread goes through three basic cycles:

1. The normal cycle before road contact;
2. The contraction cycle as it makes contact; and,
3. The expansion cycle as that section of tread leaves the road surface.

This contraction of the tread design which is illustrated in Figure 2 is referred to as squirming in the tire industry. The tire tread has distorted to conform to the flat road surface. This has caused the ribs to squeeze together and the grooves to close up a little. This squirming is more of a

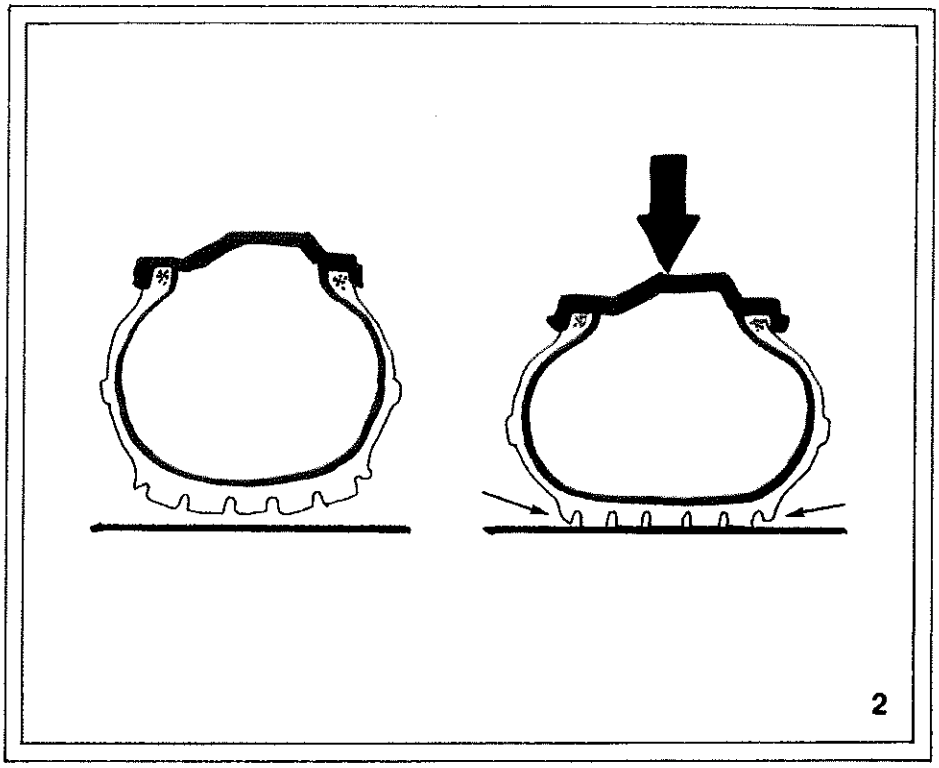
2. *Tire Tracks and Tread Marks* — Given, Nehrich and Shields, page 1.

3. *The Story of Rubber* — Firestone Canada Ltd. Publication

4. *Tread Designs: Yesterday & Today* — by Addis Finney, page 36.

problem in the bias ply tire construction and occurs to a lesser degree in the belted bias and radial ply construction tires.

In the tire industry, that area of the tread which makes contact when the tire is under load, and the impression it leaves, is referred to as the *tire footprint.* As each tread element goes through the footprint in the contraction and expansion cycle it actually vibrates producing a tone or hum. It is the squirming of the tread elements or ribs which produces the sound. Smaller elements will vibrate faster than will a larger element, hence, they will produce a different tone. If all the elements were the same size they would produce the same tone which would reach unacceptable noise levels.



2

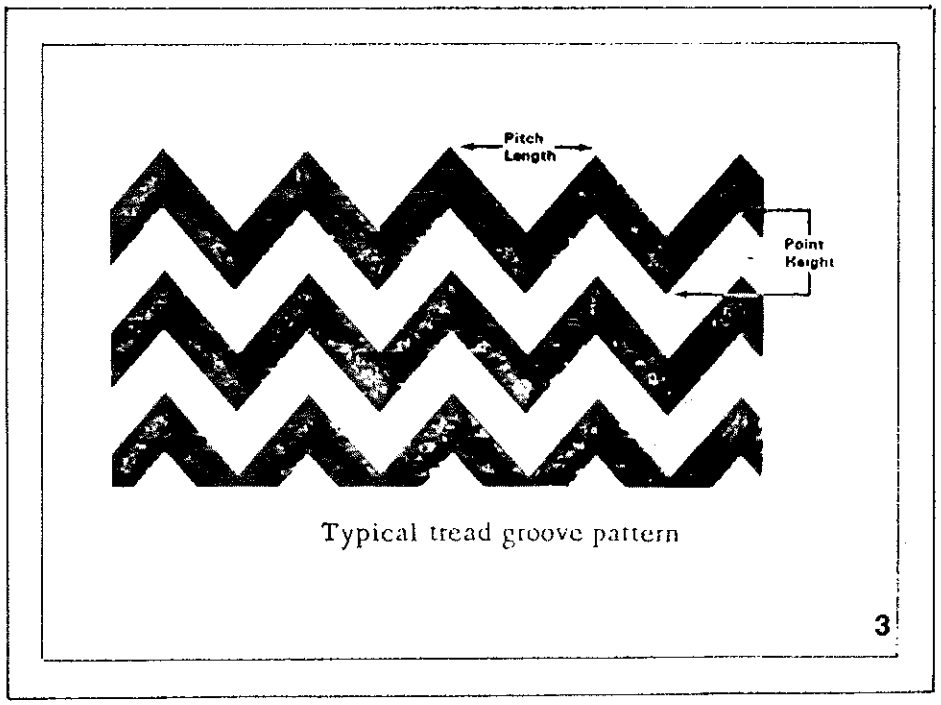
> Different sized elements vibrate at different speeds and produce different tones. Since several different tones vibrating at once can cancel each other out, tire hum is eliminated by making the tread elements in several different sizes.[5]

By arranging these different pitch lengths around the circumference of the tire in various combinations the design engineer looks for the optimum noise treatment. A quote from an article by The Firestone Tire and Rubber Co. provides clarification of pitch length.

> Point Height — Pitch Length Relationship — Another important variable in the tread design is the groove configuration. Most of the tires in the field today contain grooves which have a zigzag appearance or, in more technical terms, contain a specific point height and pitch length.[6]

Figure 3 illustrates these terms.

Today the designing process has become very sophisticated with the use of computer generated pitch sequences. In Figure 4 (Courtesy of The Firestone Tire and Rubber Company) we see Mr. Bill Wallet, manager



Typical tread groove pattern

3

5. *Tire Performance and Construction* — Resources Development Corporation, page

6. *Factors Affecting Passenger Tire Traction on the Wet Road* — J.D. Kelley, Jr., The Firestone Tire and Rubber Co., page 582.

of tire design for Firestone, examining a tread design pitch sequence on a computer screen.

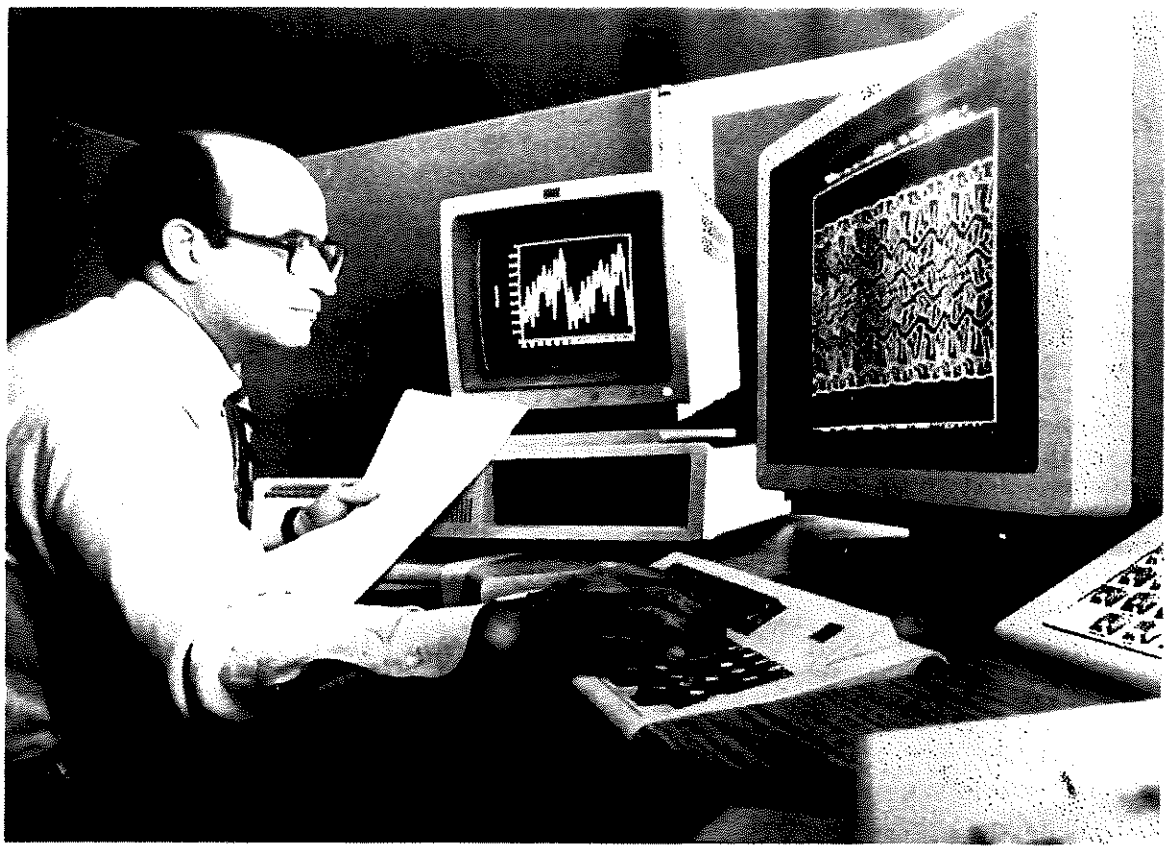
In one of their articles the BF Goodrich Tire Company gives this explanation:

> A tire designer begins this tread design process by creating a series of geometric tread-block shapes that are keyed into the computer. The computer scales and models these shapes into a complete tire tread pattern. This pattern's physical makeup is translated into computer language so the computer can predict and estimate certain performance characteristics. The engineer evaluates the tread's traction capabilities by examining the computer-generated, tread-void ratio (area of rubber vs grooves on the road), which is a determinant factor in wet and dry traction. Tire-emitted road noise is evaluated through computer analysis of tread block shapes and their pitch sequencing.[7]

All passenger tires manufactured today use some form of noise treatment or pitch sequencing, as it is also called, to reduce tire-emitted road noise. About the only tires which might not use a noise treatment would be large industrial tires which are used at low speeds so that noise is not a factor. Noise treatment is not something new to the tire industry and, in fact, has been in use since around 1930.

For the purposes of examining how noise treatment is relevant to forensic tire identification, we will be using the Firestone Super 125 P225/70R15 radial tire in this article. In Figure 3 we referred to the term pitch length in describing the groove pattern. Now look at Figure 5b (on page 6) which is a test impression taken from the Firestone Super 125. You will note that the pitch length changes from the smallest number one on the left and gets progressively larger up to number eight then reduces in unit length again progressively down to number one. This change in pitch length affects the size of the elements as well. If you examine the elements in the intermediate rib you will see they also increase and decrease in size and shape throughout the length of noise treatment. On this particular tire the pattern of noise treatment completes

7. *Research and Development T/A High Tech Radials* — BF Goodrich Tire Co.



4

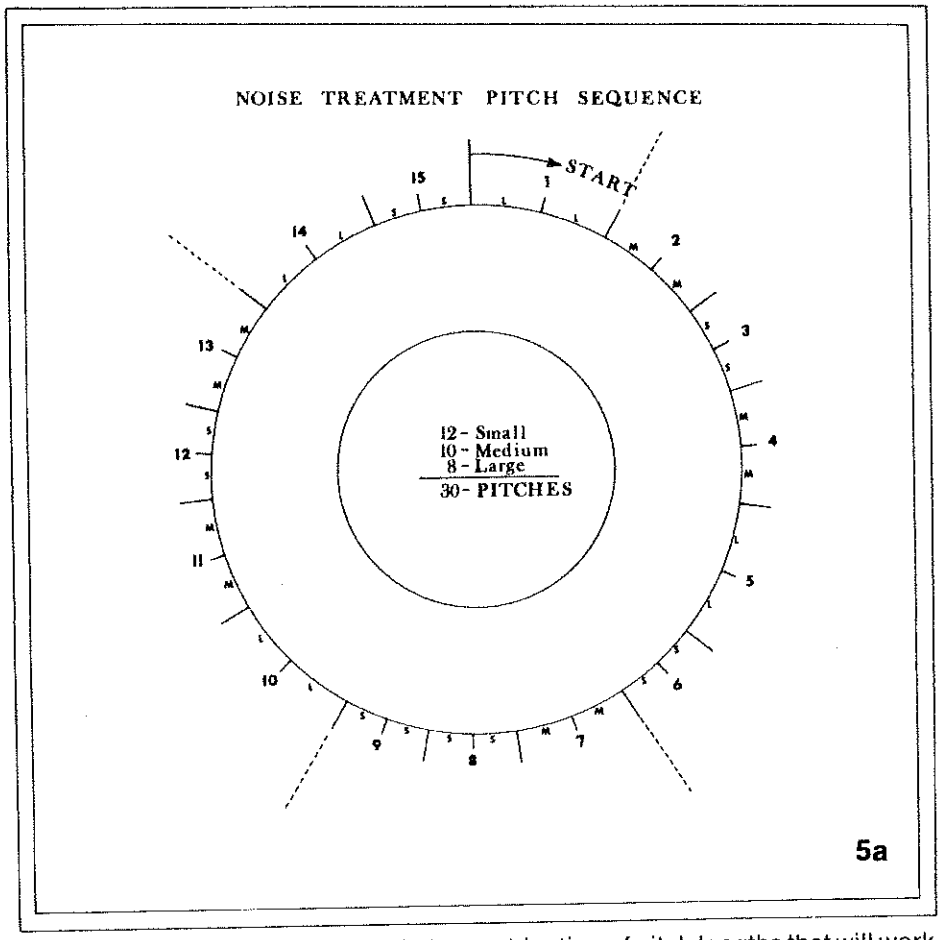
one quarter of the tire circumference. The pattern will therefore repeat itself four times for the complete tire circumference.

The noise treatment for some tires may be similar to this tire in that they increase and decrease in progressive increments. The element shapes may be different and the pitch sequence may appear as 1, 2, 3, 4, 5, 5, 4, 3, 2, 1.

The pitch lengths may, however, be arranged in any sequence which best suits the particular design. For example, the design engineer may decide on three sizes of pitch lengths — large, medium and small. These three unit lengths can be scrambled in many different arrangements to achieve noise reduction. That may happen to be LL, MMM, SSSS, MM, completing 1/3 of the tire circumference. This same sequence would then repeat twice more to make up the entire tread design. If a tire with such a design made a crime scene impression and the crime scene impression was subsequently compared to a full circumference test impression from the recovered tire, an agreement in noise treatment could be established in three locations on the test impression. This comparison process will be explained in more detail later on in this article.

Another example using three sizes of pitch lengths — large, medium and small is given in Figure 5a. In this example the pitch lengths are arranged in pairs — LL, MM, SS around the tire circumference. To complete the tire circumference 30 pitch lengths were used, 12 small, 10 medium, and 8 large. The pairs of pitch lengths have been numbered 1 to 15 for reference purposes only. If you examine the pitch sequence from sections 1 to 6 and 10 to 15 you will note that they occur in the same order. If a crime scene impression was left by either of these areas and then compared to a test impression from the offending tire it would agree in two locations on the test impression.

Now examine the pitch sequence in sections 7, 8, 9 and 14, 15, 1. These



areas contain an arrangement of pitch lengths which are not repeated in any other locations on the tire circumference. If you are dealing with these areas of tread design then you will be able to locate the exact place in the test impression from where the crime scene impression came.

As well, you will find that when dealing with very short sections of tread design it will tend to fit into more locations in the test impression, because you are not working with enough pitch lengths. Generally speaking the longer the sections of tread design you are working with the better you will be able to locate the area in the test impression from where it came.

It will become apparent if you examine test impressions from various tires that noise treatments come in many forms. Design engineers refer to these arrangements of pitch sequences as arithmetic and logarithmic noise treatments. Tire tread designs are man-made creations and virtually any combination of pitch lengths that will work are possible.

The number of times noise treatment patterns repeat themselves on a tire circumference will vary for different tread designs. Some designs call for patterns that repeat two, three or more times. I have worked on cases where the pattern repeats several times and then one sequence of pitch lengths is not found anywhere else on the tire circumference, as pointed out in Figure 5a. With the use of computers, some tread design noise treatments do not repeat at all and, therefore, have no two locations on the circumference with the same arrangement of pitch lengths.

The length of the noise treatment pattern in Figure 5b which makes up ¼ of the circumference is 576 mm. This is a tread design for a 15 inch rim. Firestone maintains the same noise treatment for different size tires by making the treatment proportionately larger or smaller to accommodate the different circumferences. This means the same tread design in a 14 inch rim

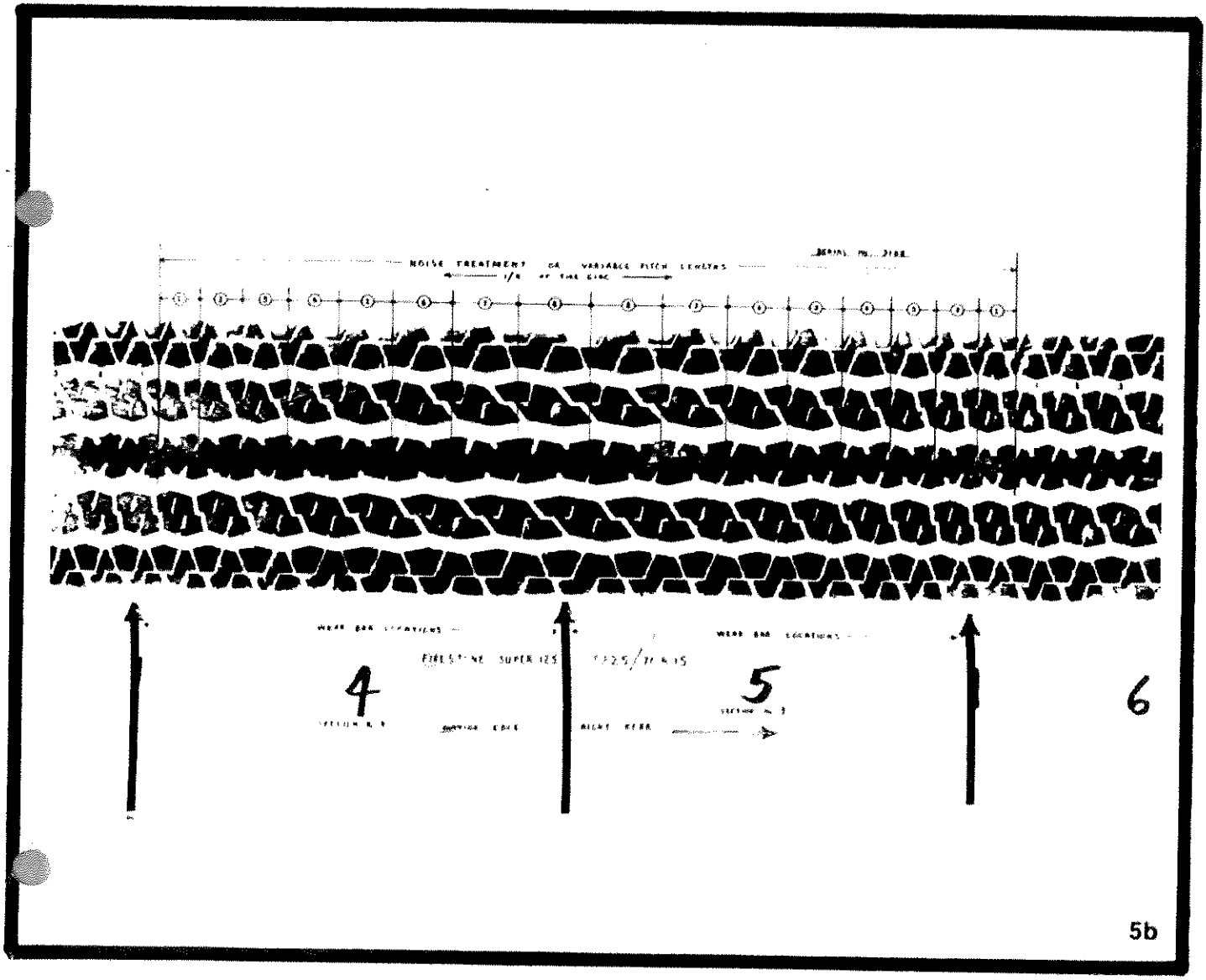
size would have a shorter length of noise treatment to fit on the smaller tire circumference. Most tire companies make their noise treatments proportionately larger or smaller to fit the different tire sizes. This can be a very useful aid in the comparison process to differentiate tires with the same tread design. There are some tires made that maintain the same length of noise treatment and add extra unit lengths (pitch lengths) to make up for the larger circumference. Moving up in size, however, generally increases the width of the tire which will reflect itself in a change to the noise treatment.

WEAR BAR INDICATORS

Also of importance in tire impression identification are wear bar indicators and their relationship to the noise treatment on the tire. Wear bars are located in the grooves and run laterally across the tread. (See Figure 1.) They are raised one sixteenth of an inch above the base of the groove. When the life of the tread has expired, wear bars show up as a bald strip across the face of the tread. Wear bar indicators may therefore show themselves in crime scene impressions as illustrated in Figure 6. Where the crime scene impression lends itself to casting, the three dimensional cast may also record wear bars.

As can be observed in Figure 7, the noise treatment pattern repeats four times around the circumference and the wear bar indicators are located at eight intervals. The Firestone Super 125 presents a particular relationship in which wear bars always appear in the same location with respect to noise treatment in each quarter section. As we have already mentioned, however, the number of times the noise treatment repeats itself will vary for different designs. Likewise the number of wear bars may vary. On tires of less than three hundred and five millimeters/twelve inches, there are at least three indicators. Tires with a rim diameter of three hundred and five millimeters/twelve inches, or more, have at least six indicators.

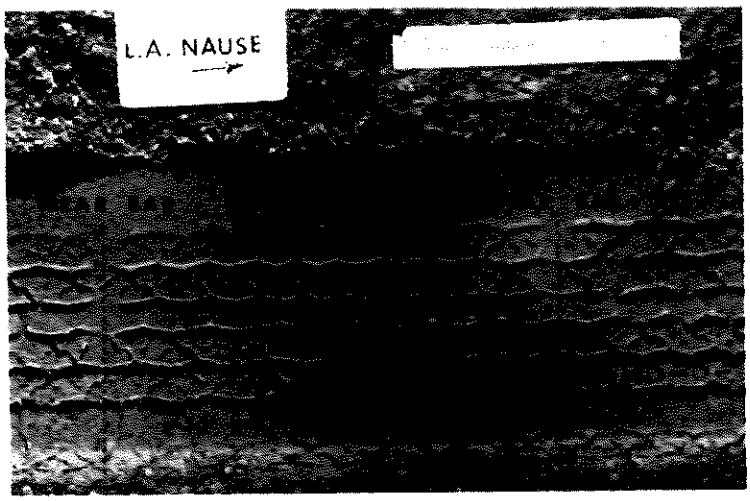
The major tire companies use a precision cast method for making their aluminum alloy molds. This involves



5b

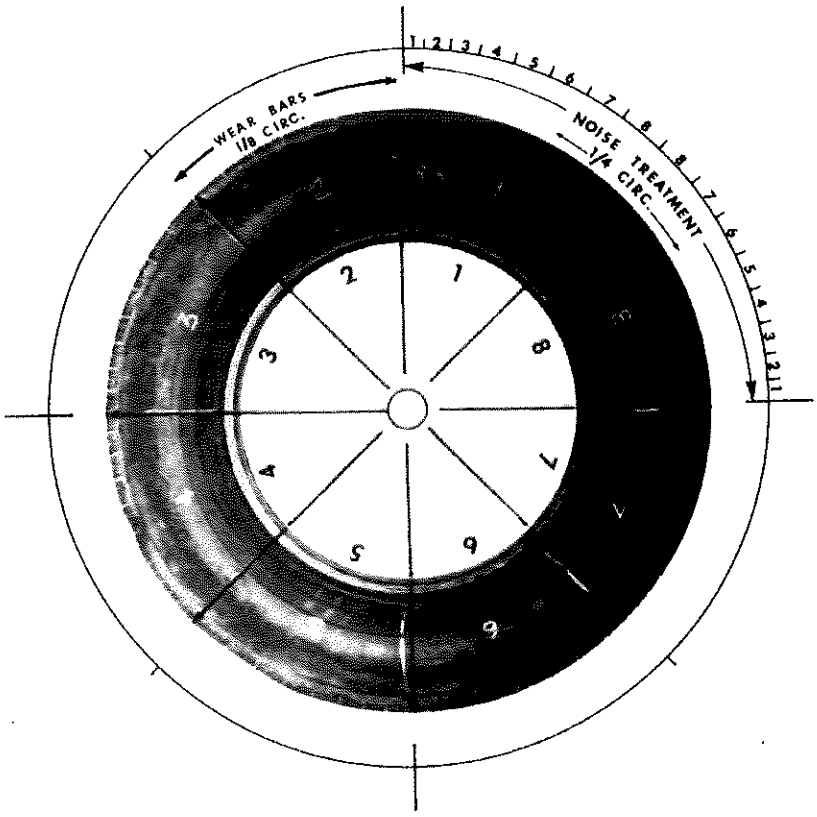
the use of plaster cast positives made from master models to prepare any number of identical tire molds. Using this method the wear bars are engraved into the master models and therefore are located in the same place on all tire molds produced from these masters.

For companies producing a line of tires which will have a lesser volume of sales it would be too costly to use the precision cast method. They may only require two or three tire molds of a certain type. For this reason they may prefer to have each aluminum alloy mold individually engraved. Each individual mold in this case may have the wear bars placed in different locations respecting the noise treatment, as well there may be slight differences in the noise treatment itself.



6

NOISE TREATMENT/WEAR BAR RELATIONSHIP



7

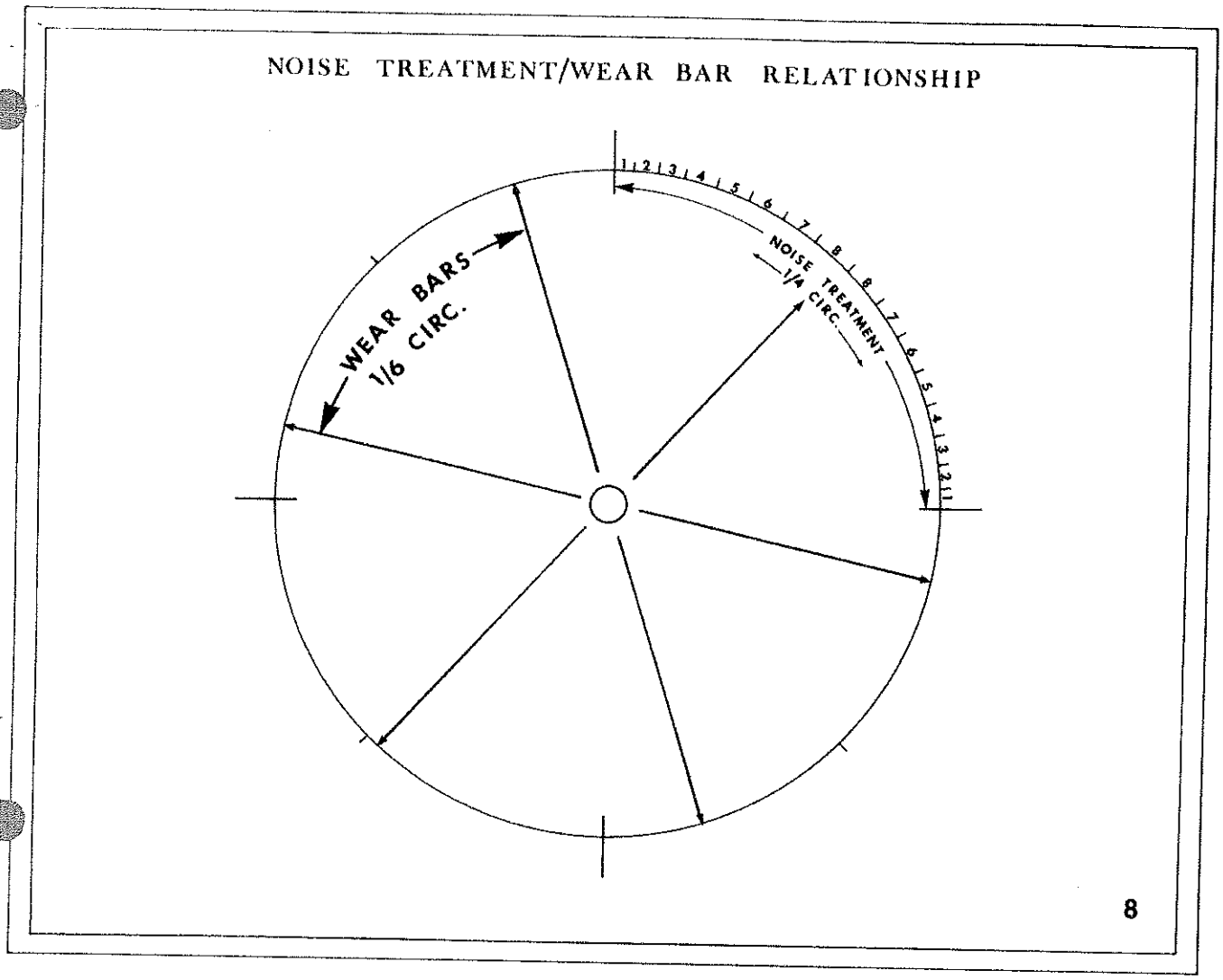


Figure 8 shows another variation of the relationship of wear bars to noise treatment. In this case there are only two locations that have the same wear bar — noise treatment relationship.

Because of these variations, if wear bars show up in crime scene impressions, they can help locate the area of noise treatment that made the particular impression. In photographing crime scene impressions and making plaster casts it is a good practice to try and include 600 millimeters/24 inches of impression if possible. This will generally record the location of two wear bars if they are present. By measuring the distance between the wear bars, you will have another piece of comparison information. In most cases, wear bars are equally spaced around the circumference, the larger the rim size the further apart the wear bars will have to be spaced. On most tires wear bar indicators are quite easy to observe in the grooves. In some of the more intricate block shaped tread patterns the wear bars blend in very well and can be more difficult to spot. They will generally not be registered in crime scene impressions unless the tread is worn down to their level.

It should be pointed out that the location of wear bar indicators in a crime scene impression is a class characteristic, since they are placed there by the manufacturer and will be in the same location on tires coming from the same molds. Their presence in crime scene and test impressions does, however, indicate an agreement in tread wear.

TIRE IMPRESSION COMPARISON AND IDENTIFICATION

The comparison and identification of tire impressions employs many of the same techniques and principles involved in the identification of footwear.

When comparing crime scene and known tire impressions the examination may be broken down into three basic steps of the scientific approach;

Analysis The determination that the crime scene and test impressions are tire impressions with similar class characteristics and thus warrant closer examination.

Comparison The impressions are then examined more closely for any agreement of class and accidental characteristics which may be present.

Evaluation Consideration is then given to the degree of agreement or disagreement of the class and accidental characteristics in order to arrive at an opinion.

By following these three basic steps the examiner will have covered the following areas of comparison:

1. Examined the overall class characteristics to determine if the crime scene impression and suspect tire are of the same tread design. (Note: It may be possible to find the same tread design on different brand name tires. Some of the major tire manufacturers make tires for other companies and may use the same tread design changing only the sidewall brand name.);

2. Checked to ensure that the crime scene impression is the same size as the suspect tire;

3. Searched for one or more locations on the suspect tire impression that could have made the crime scene impression;

4. Checked for any variations between crime scene and test impressions;

5. Examined the crime scene and test impression for agreement or disagreement of tread design wear; and,

6. Searched for the agreement of any accidental characteristics present.

It is the quality and number of accidental characteristics in agreement which enables the specialist to make a positive identification of crime scene impression to suspect tire. There is no set number of accidental characteristics required to make a positive identification. The reason for this is that several factors enter into consideration before an opinion is expressed — the examiner's experience, the uniqueness of the accidental characteristic and the clarity with which the accidental characteristic is reproduced in the crime scene impression.

Each case must be judged on its own merits. Care must be taken in expressing any form of opinion evidence. Nothing will ruin your credibility more quickly than the tendency to overstate the value of your evidence.

The *opinion* may be expressed in one of the following ways:

1. The crime scene tire impression is identified as having been made by the tire in question and that only the tire in question could have made the crime scene impression;

2. The chances of another tire having the same agreement of class and accidental characteristics is remote or unlikely;

3. The crime scene tire impression is consistant with having been made by the suspect tire or any other tire of the same tread design and size; and,

4. The crime scene tire impression is not consistant with having been made by the suspect tire in question.

Due to the fact that there may be four different tread designs located in four different positions on the vehicle, the evidence can sometimes be quite incriminating even in the absence of a positive tire identification. This enters the field of forensic vehicle identification, which among other things also considers vehicle dimensions recorded at the crime scene. This was discussed in more detail in my previous article "Tire Impressions as Evidence" (*RCMP Gazette*, Vol. 44, No. 12, 1982).

DEFINITION OF A TIRE FOOTPRINT IDENTIFICATION EXPERT

Early on in my research of this subject, I was put in touch with Mr. Peter MacDonald who was then manager of tire design for the Firestone Tire and Rubber Co. in Akron, Ohio. At that time he was good enough to answer many questions and provide me with some much needed information. Mr. MacDonald, at that time, had assisted some police forces on actual investigations involving tire impressions and continues to provide assistance in this area as well as lecturing on the subject.

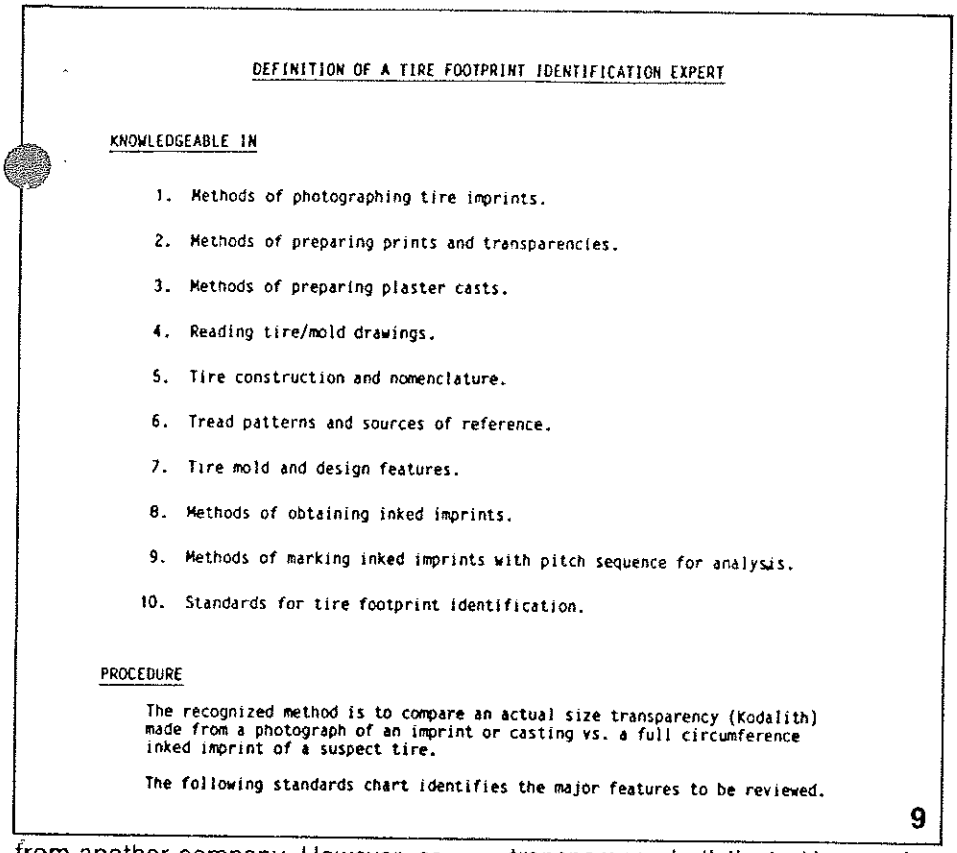
During my recent tour of the world headquarters facilities for Firestone and BF Goodrich in Akron, Ohio, I had the opportunity to spend some time with Mr. MacDonald and discuss at length the science of tire impression identification. He has now retired from Firestone and opened a consulting firm called Tire Forensics. His experience in designing tire treads for so many years and in working with several police forces on investigations dealing with tire evidence has given him an excellent understanding of forensic tire impression identificatior

We are both in agreement on the fact that more information and training should be made available to specialists dealing with this type of evidence.

I thought you might appreciate seeing two lists which Mr. MacDonald has prepared to assist those working in the field of forensic tire impression identification. The lists are meant to be used as guidelines to follow and are not intended to be interpreted as hard and fast rules.

Definition of a tire footprint identification expert is given in Figure 9. These are all points which I feel would be useful to the identification expert, however, not all are necessarily a requirement for becoming an expert in this area. Most of the points are self-explanatory, however, there are two points, number four and number nine which I would like to discuss briefly.

Reading tire/mold drawings (point number four) are, I feel, useful to have if they are available. I have had the opportunity to examine tire mold drawings and study how they relate to a tire test impression. When you produce a test impression from a tire you are also reproducing the design and dimensions used to make the tire mold for that tire. When you obtain the mold drawings it is like having the key to unlock a puzzle. It will provide you with the arrangement of pitch sequences used by the tire designer for that tire's noise treatment. Without the mold drawings the pitch sequencing is very difficult to figure out even for a design engineer



DEFINITION OF A TIRE FOOTPRINT IDENTIFICATION EXPERT

KNOWLEDGEABLE IN

1. Methods of photographing tire imprints.
2. Methods of preparing prints and transparencies.
3. Methods of preparing plaster casts.
4. Reading tire/mold drawings.
5. Tire construction and nomenclature.
6. Tread patterns and sources of reference.
7. Tire mold and design features.
8. Methods of obtaining inked imprints.
9. Methods of marking inked imprints with pitch sequence for analysis.
10. Standards for tire footprint identification.

PROCEDURE

The recognized method is to compare an actual size transparency (Kodalith) made from a photograph of an imprint or casting vs. a full circumference inked imprint of a suspect tire.

The following standards chart identifies the major features to be reviewed.

9

from another company. However, as you will see later on in this article, it is not necessary to know what the pitch sequencing designation is for a tire test impression in order to locate the area on the test which could have made the crime scene impression. Also, some tire companies would refuse to supply such classified and restricted drawings, so I feel one need not be overly concerned if mold drawings are unfamiliar to you or unavailable for a specific case.

Methods of marking inked imprints with pitch sequence for analysis (point number nine) is along the same lines as point number four. If you don't have the mold drawings it is very difficult to mark the inked impressions with the same pitch sequencing used by the manufacturer. It is important to be aware of the fact that the pitch lengths do change as you go around the tire and that the crime scene impression could only have come from certain locations on the circumference. As for the actual pitch sequencing formula used by the manufacturer, that is not critical to the comparison. Once areas have been located on the suspect test impression, using a full size Kodalith transparency, both the test impression and Kodalith can be marked in your own way in order to quickly locate these areas again for comparison purposes.

TIRE FOOTPRINT IDENTIFICATION STANDARDS

The tire footprint standards are listed in Figure 10a. This is a very detailed list prepared by Mr. MacDonald to assist tire identification experts in the comparison and identification process.

Since you may not be familiar with the purpose of each point listed I will take the liberty of elaborating briefly on them.

CLASS CHARACTERISTICS

These are features of the tread design created during the manufacturing process which are common to all tires of that same size and design.

The class characteristics are sub/divided into categories A and B.

Category A — Brand vs Brand

This compares one brand of tire with another brand or even tires of the same brand name, but of different sizes, according to the following characteristics:

1. *Element shape* — comparing element shapes on one brand of tire versus a similar tread design, which has similar shaped elements.

2. *Number of ribs* — rib count on the tread design. Includes a look at the width of ribs as well.

3. *Groove shape* — sometimes the groove shape stands out better than the rib or element shape. This may be easier to look at in comparing two impressions or in searching through the *Tread Design Guide* in order to come up with a brand name for crime scene tire impressions.

4. *Sipe pattern* — sipes are thin grooves in the ribs and elements. Close examination of the sipe pattern can help distinguish between tires of similar tread designs or element shapes.

5. *Noise treatment* — comparing the noise treatment of a crime scene impression using a full size Kodalith transparency can determine if the tread is the same design and size tire, and how many places on the tire circumference could have made the impression.

6. *Arc Width* — the width of the tread design. Sometimes difficult to measure on tires with rounded shoulders. As a tire tread wears down it can cause the measureable arc width to increase slightly on designs which have rounded or sloping shoulders. See Figure 10b (on page 12).

7. *Notches* — a groove which enters the side of a rib or element but does not continue on through. Check ribs and elements to see that notches are in agreement with crime scene impression.

8. *Slots* — a groove which runs laterally across the tread from one circumferential groove to the next. Check for agreement between the crime scene impression and the suspect tire. Like notches they

can also change in appearance with tread wear.

9. *% Void* — look for agreement in the area of rubber vs groove space.

10. *Stud pattern* — some tires of same brand name and design may or may not have stud holes in m/s tires. The stud holes in crime scene impressions should be in agreement with area located in test impressions.

11. *Side treatment* — this is the design on the shoulder of the tire which improves traction and also the look of the tire. From time to time it shows up in crime scene impressions.

12. *Round shoulder vs square shoulder* — this feature can sometimes be quite distinctive in a crime scene photograph or plaster cast and help rule out certain suspect tires.

13. *Blackwall vs whitewall* — some tread designs and widths only come with whitewalls or raised white letters on the sidewall of the tire. This may be useful to the investigator in keeping an eye out for possible suspect vehicles.

Basically when a comparison is made for agreement of noise treatment, as was pointed out in item number 5b, it will automatically cover many of the points just listed. As well these points listed under brand vs brand are not only useful in comparing crime scene impressions to suspect tire impressions, but are also useful in searching a crime scene impression through the *Tread Design Guide* in order to come up with a brand name and photographs for the type of tire which made the crime scene impression. The *Tread Design Guide* and *Who Makes It and Where* are both useful reference sources of tire tread designs published every year by Tire Guide, The Tire Information Center, P.O. Box 677, 14 Jackson Avenue, Syosset, N.Y. 11791.

Category B — Mold vs Mold compares tires of the same tread design, which may have discrepancies imparted by different molds for the following:

1. *Mold rotation* — full circle molds which have a top and bottom half that close on the green tire in the vulcanization process may not meet exactly as intended. The top and bottom halves may be slightly offset and need readjustment. The full circle mold produces a seam around the circumference of the tread design where the two halves met when in the closed position. This offset can sometimes be observed

TIRE FOOTPRINT IDENTIFICATION STANDARDS

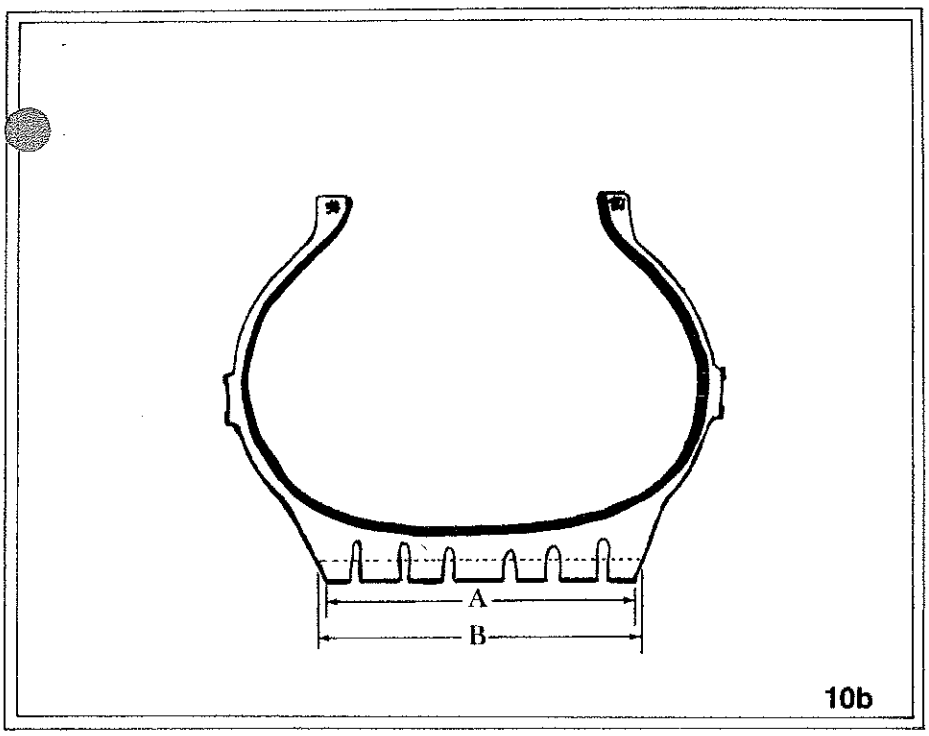
CLASS CHARACTERISTICS				ACCIDENTAL CHARACTERISTICS			
A	BRAND VS. BRAND	B	MOLD VS. MOLD	C	GENERAL	D	SPECIFIC
1	ELEMENT SHAPE	1	MOLD ROTATION	1	CIRC. WEAR	1	CUTS
2	NUMBER OF RIBS	2	TREAD WEAR INDICATORS	2	LATERAL WEAR	2	TEARS
3	GROOVE SHAPE	3	MOLD VARIATIONS	3	CUPPING	3	CHUNK OUTS
4	SIPE PATTERN	4	SERIAL SIDE IN VS. SERIAL SIDE OUT	4	HEEL AND TOE	4	STONE HOLDING
5	NOISE TREATMENT			5	SKID DEPTH	5	TEXTURE VARIATIONS
6	ARC WIDTH			6	EXPOSED TIE BARS	6	ABRASIONS
7	NOTCHES			7	FURROW WEAR	7	SIDE TREATMENT VARIATIONS
8	SLOTS						
9	% VOID						
10	STUD PATTERN						
11	SIDE TREATMENT						
12	ROUND SHO. VS. SQ.						
13	BLACK S.W. VS. WHITE						

ALL FEATURES IN SECTIONS A, B, AND C ARE TO BE REVIEWED IF APPLICABLE, FOR CORRELATION

ALL FEATURES IN SECTION D ARE TO BE REVIEWED, IF APPLICABLE, FOR CORRELATION.

IT MAY BE POSSIBLE - WITH ONLY ONE (1) SPECIFIC ACCIDENTAL CHARACTERISTIC TO MAKE A POSITIVE IDENTIFICATION. HOWEVER, MORE THAN ONE (1) SPECIFIC CHARACTERISTIC GENERALLY SHOULD BE IDENTIFIED FOR A POSITIVE IDENTIFICATION.

10a



10b

in elements that do not meet precisely in this area. See Figures 11 and 12 which have slightly different mold offsets. In checking the serial and mold numbers on the two tires it was noted they came from different molds. In good quality crime scene impressions this mold offset can be detected and help eliminate suspect tires with the same tread design. If a mold offset becomes more than a 1/10th of an inch it is usually spotted and adjusted. Minor offsets of this nature do not affect tire performance or quality. A second type of mold which is becoming more common with the advent of high quality radial tires is the segmented mold. This mold is separated into equal segments that produce seams which run laterally across the face of the tread. The segmented mold does not have mold offset problems.

2. *Tread Wear Indicators* — as already mentioned they may be used to locate specific noise treatment area on test impression which may have made crime scene impression. They may also help eliminate similar suspect tires if wear bar location or distance between them is not in agreement.

3. *Mold Variations* — the small metal plates in the tire molds which produce the sipes in the finished tire are thin and may become bent. This would then produce a tire which had distorted sipes in a particular location. Quality control in the tire industry is of a very high standard and this would be a rare occurrence and corrected as soon as it was detected.

4. *Serial Side In vs Serial Side Out* —The serial tin side is the side of the tire which has the serial number recorded on it by the serial tin plate in the bottom half of the mold. The serial number is embossed on the inner sidewall and mounted on the inner side of the rim. The outer sidewall contains such things as raised lettering and whitewall markings. In cases of directional tread designs if the tire is mounted on the rim in reverse (serial side out) it will change the direction of the impression left by the tread design. It is uncommon to have directional treads mounted in the wrong direction so this could be another variable to watch for in tire comparison.

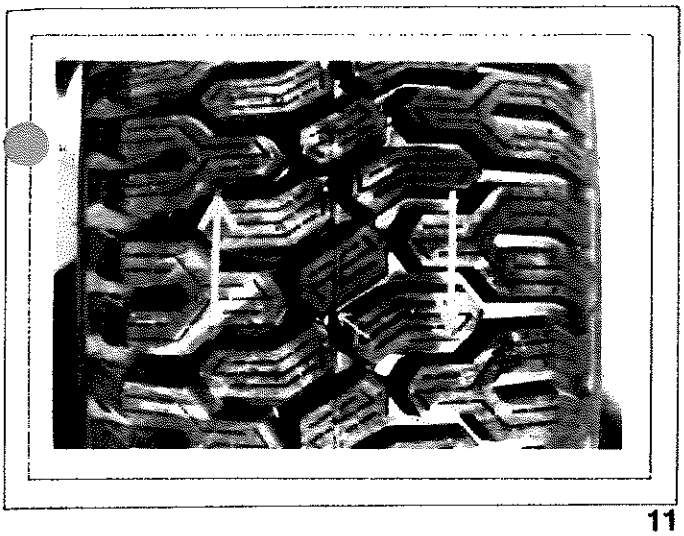
ACCIDENTAL CHARACTERISTICS

In forensic tire identification these would be cuts, tears and wear features placed on a tire as a result of being subjected to the functions it performs as a vehicle component. Accidental characteristics may also be marks unintentionally made by the manufacturer but not reproduced in subsequent tires. (Such as small nicks or cuts made in the vent trimming process.)

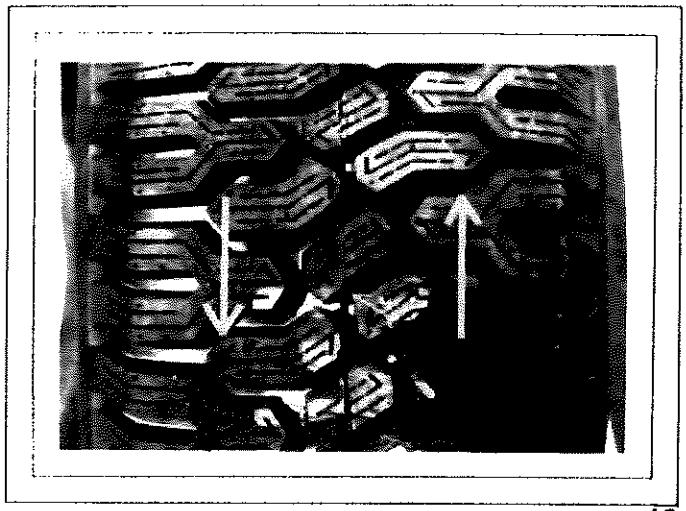
Accidental characteristics are subdivided into C and D categories:

C — General — Routine or irregular wear to the tread design as it changes throughout the life of the tire. As a tire functions as a vehicle component it is subjected to various stresses which cause the tread to wear away. Mechanical problems with the vehicle can cause accelerated or irregular tire wear. The squirming action of the tread elements as they make road contact also scrubs away tread rubber. The first several thousand miles produce the fastest tread wear. As the tread elements become shorter, they are less flexible and squirming is reduced, so the tread wear slows down considerably. One could expect a small accidental characteristic to last longer on a well worn tire than on a new one, all other things being equal.

1. *Circular Wear* This is a description of how the tire is worn around the circumference of the tire with sipes and grooves indicating amount of wear. As a new tire tread wears away it can often change its appearance from the basic design. Most tires being made today use siping of two or three different depths. Some of the sipes will go all the way to the bottom of the pattern depth, while others will only go down part way. The solid rubber under the shallower sipes is referred to as tie bars, see Figure 13. The tie bars are necessary to hold the tread elements rigid so they do not squirm too much. Less squirm saves on heat buildup and tread wear life. When the tread wears down to the tie bars, it changes sipe appearance. Consequently there is less siping on a worn tire than on a new tire, as a rule. This is something



11



12

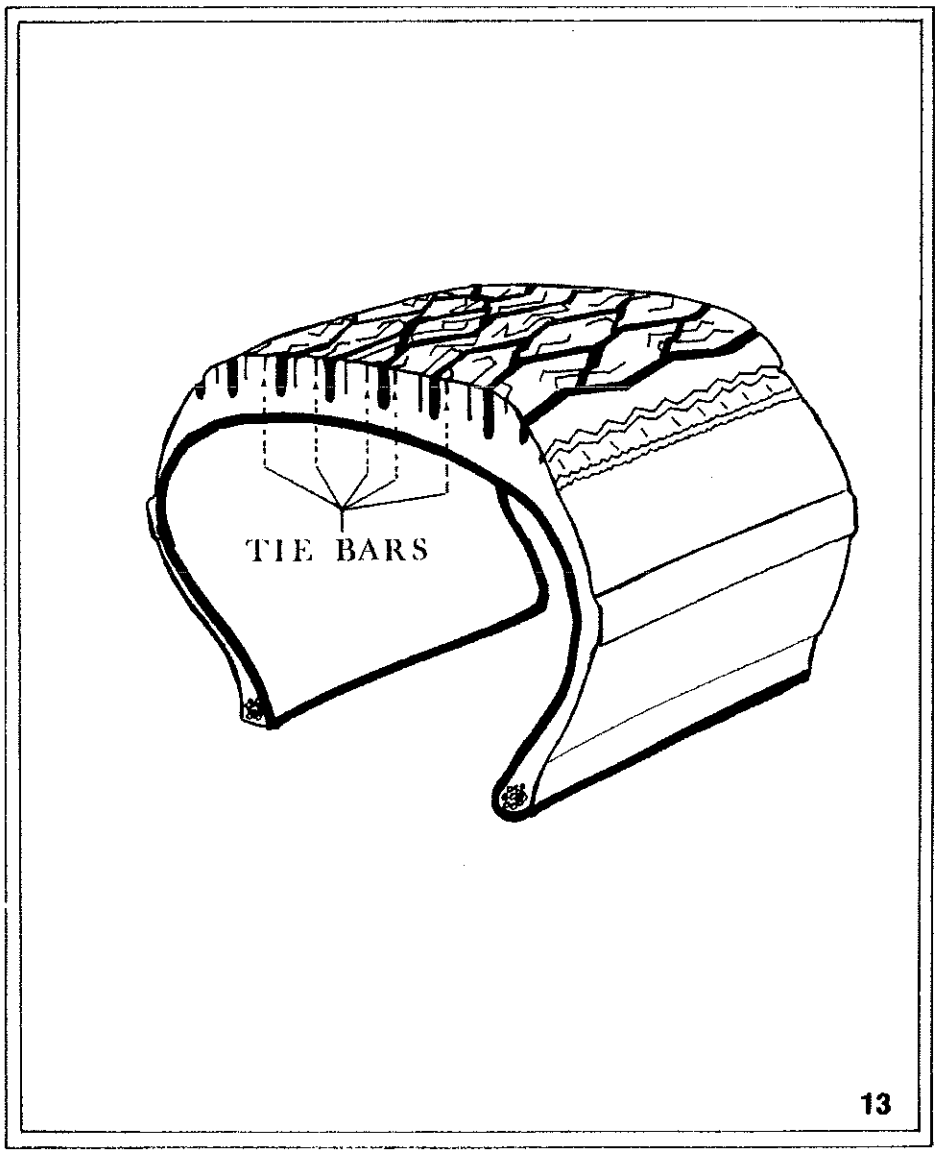
which is of concern to tire manufacturers because it does not create its own problems.

> One criticism of present day siping practice might be that too many are used for styling considerations, and that they are made less than full depth in part of their length in order to maintain tread element stability. As a result, the siping pattern partly or completely disappears during the latter part of the tread life, just when the shallower design depth makes sipes most needed.[8]

Some of today's high performance tires are using designs with well-defined, tread-block shapes that do not have sipes in the design at all. Instead these designs rely on grooves, lateral slots and notches to prevent hydroplaning and improve traction.

Examine Figure 14 (on page 14) to see how the sipes and notch design has changed in new and worn tires of the same design. This is an important factor to consider when doing comparisons. It may help eliminate a similar tire as not being responsible for the crime scene impression.

2. *Lateral wear* — tread wear from a cross-sectional perspective.
3. *Cupping wear* — depressions or a cupping effect in the tread design caused by some excessively loose suspension part which allows the wheel to oscillate according to a frequency pattern. The location of



13

8. *Rubber World — Tread Designs: Yesterday & Today* by Addis Finney.

the depressions would occur randomly in the noise treatment around the tire circumference.

Heel and Toe — wearing down of ribs or elements so that they are lower on one side than they are on the other. The low wear end is referred to as the heel and the higher part as the toe. This is the type of wear you might expect to see on the front tires of a vehicle which has incorrect toe-in or toe-out settings.

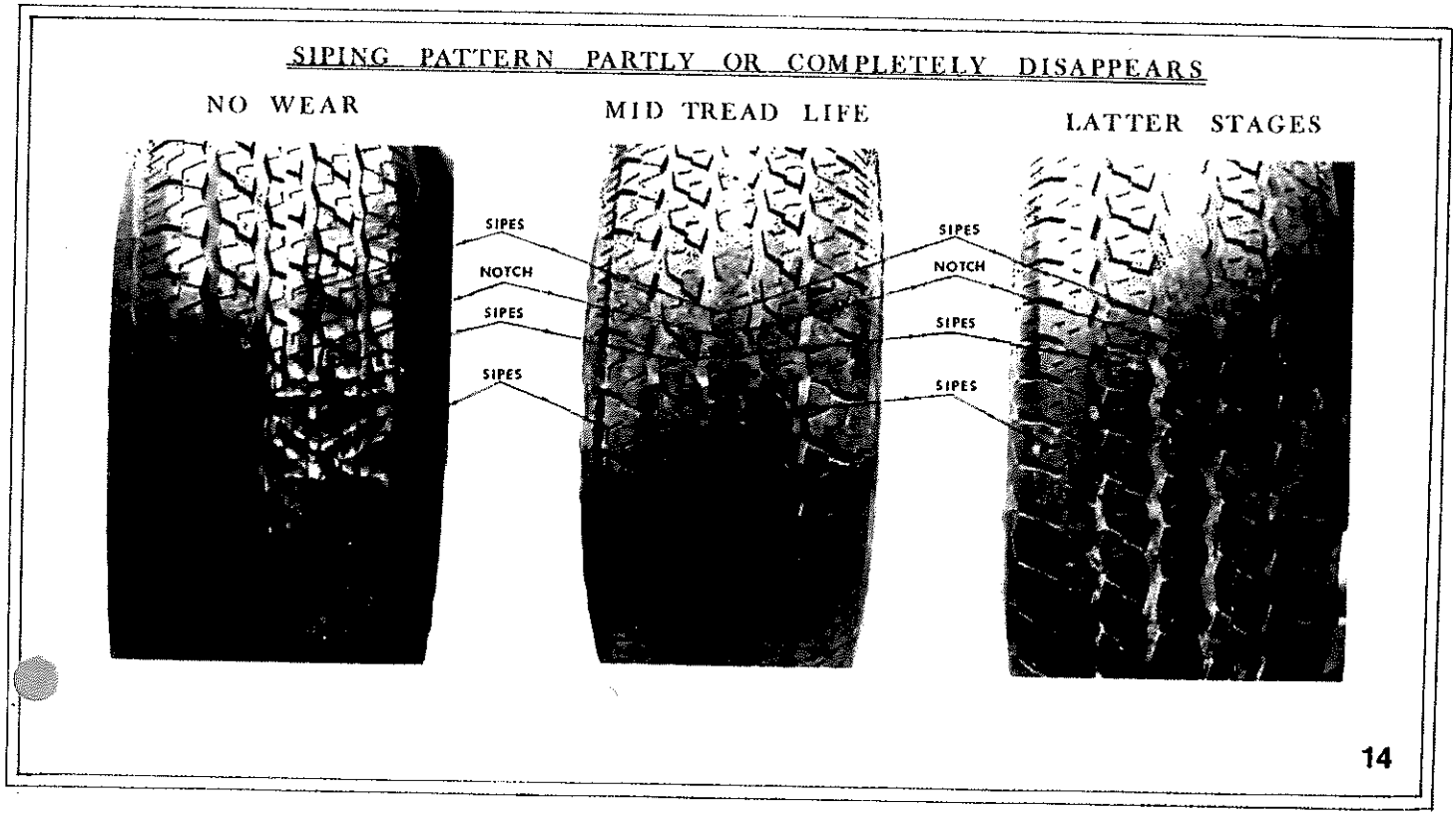
5. *Skid Depth* — The amount of measureable tread left on the suspect tire as compared to that recorded in the crime scene impression. Most of the tread designs today start as a depth of 11/32 to 12/32 of an inch. Mud and Snow (M/S) type designs mostly start at around 14/32 to 16/32 of an inch. The recognized measurement for tread depth in North America is recorded in 1/32" intervals with the use of a tread depth gauge. The gauge is inexpensive and can be purchased at most auto supply outlets. This device can be used at the crime scene to measure tread depth in three dimensional impressions where possible as well as recording the tread depth on suspect tires.

You are not necessarily trying to record the same measurement between crime scene and suspect tire, but whether the suspect tire could have made the crime scene impression. For example, if the recorded tread depth at the scene was 8/32" and the suspect tire tread depth was 10/32" then it could have made the impression without recording its full tread depth. On the other hand if the suspect tire tread depth was only 4/32" it could not have made the crime scene impression.

When the tread depth is worn to 2/32" the wear bar indicators will begin to appear as a solid bar of rubber across the tread.

6. *Exposed Tie Bars* — As mentioned in circular wear, when the tread design wears down the tie bars in the sipes, notches and lateral slots become exposed in varying amounts around the circumference of the tire. This changes the appearance of these features in the design.

7. *Furrow Wear* — Sometimes referred to as erosion wear and more commonly seen in truck tires. Because grooves do not run in a straight line and tend to zigzag there are sharp points along the groove where it changes direction. When the tire is rolling the points tend to bend and tuck into the grooves because they are less supported. They deflect and then spring back after coming out of the compression cycle. Since they bend into the groove they don't wear as quickly as the stablized areas which make solid road contact. This causes irregular furrow wear along the groove around the circumference of the tire.



Category D — Specific Definite and distinctly formulated marks of wear which in sufficient number and quality make the tire unique.

1. *Cuts* — This would refer to cuts inflicted to the tread design by sharp objects such as rocks, glass, etc. A tire tread will cut much easier when it is wet since water tends to act as a lubricant. Often stones become lodged in slots, notches or sipes which can tend to drill themselves into the tire during the contraction and expansion cycle. This can cause cuts and tears to appear frequently in these areas.

2. *Tears* — a rip in a rib or element running across the surface of the tread.

3. *Chunk outs* — tread elements can become excessively hardened by ozone in the atmosphere and do not flex as easily as they should. This may cause chunks of tread rubber to be torn off when they come in contact with sharp or rough surfaces. Excessive tread-element movement can also cause "chunk outs" by heat buildup and a weakening at the base of the elements.

4. *Stone Holding* — stones picked up and held in the tread design which are lodged in specific locations in the noise treatment.

5. *Texture Variations* — scratches or striations on the surface of the tread.

6. *Abrasions* — tears or large cuts in the shoulder or sidewall area of the tire.

7. *Side Treatment Variations* — any odd shaped minor cuts, scratches etc., to the shoulder or sidewall area of the tire.

PRACTICAL APPLICATION TO CRIME SCENE IMPRESSIONS

Now lets take a look at how this theory applies in practice with the same tire used in Figures 1, 5, and 7.

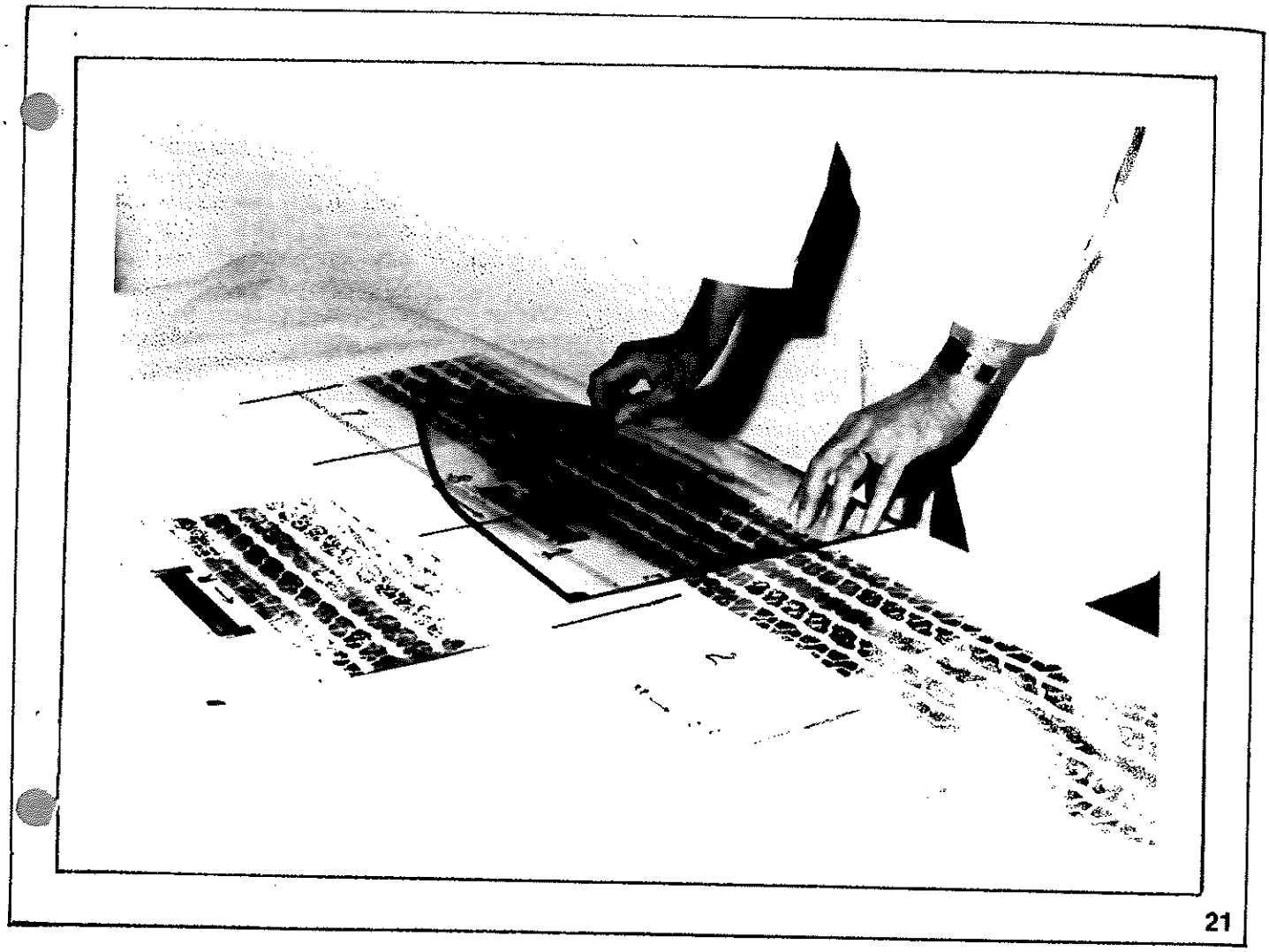
15

Tire impressions left at the scene of a crime come in many shapes and forms. For the purposes of this case study I have used a two dimensional dust impression recorded on a flat surface. The procedures and techniques of comparison used in this case study, will also work equally well for other forms of tire impression evidence, the success of which will depend on the quality of those impressions.

The tire was mounted on a vehicle and driven over paved roads and then parked in a paved parking lot with no other special attention given to it.

The vehicle was then driven over a piece of cardboard box (as seen in Figure 15) obtained from a local store. The exhibit was then recovered. If the direction of travel and tire location is known for the vehicle it should be marked on the exhibit. In this case from evidence at the scene if I knew it was the right rear tire that had made the impression I would mark it as such. As well, if direction of travel can be determined I place an arrow on the exhibit or in the photograph pointing to the front of the vehicle. This arrow acts like the laterality letter R in fingerprint photography. Most tread designs are non-directional and when reversed will appear the same, since one half of the tread design is like a mirror image of the other. By having the arrow in the photograph or on the exhibit you will always know which is the inside and outside edge of the tire impression. Marking the impression by placing the letter R or designation of tire position will also help maintain laterality in the darkroom, but this is over and above some method of marking for inside and outside edge of the impression, if that fact is known.

This type of impression is not easily observed since it is light colored on a light surface. It can be seen from an oblique angle, however, it is difficult to photograph. Using a technique which is well known in footwear identification the dust impression was lifted using the carbon paper method. This involves securing the exhibit from moving and covering it with a 14" x 17" sheet of carbon paper. The carbon paper is then held from slipping while it is stroked with a piece of fur or like substance to create static. The carbon paper was then removed and contained a reversed lift of the dust impression. The arrow to maintain direction to the front of the vehicle was transferred to the carbon paper using a grease marker. A reversed letter R was then marked on the carbon paper so when



21

excellent dimensional stability. In tests conducted to measure tire footprint geometric characteristics C.E. Prettyman reported this concerning a size FR78-14 steel belted radial tire which he examined.

> As the tire is loaded the footprint at first grows in two dimensions; but after the full footprint width is obtained at about 300 lb load, further growth is achieved only along the length of the footprint.[9]

So if the *suspect tire* arch width is narrower than crime scene impression arch width by a measurable amount, say ½ inch, and no explanation can be found, the tire may be ruled out.

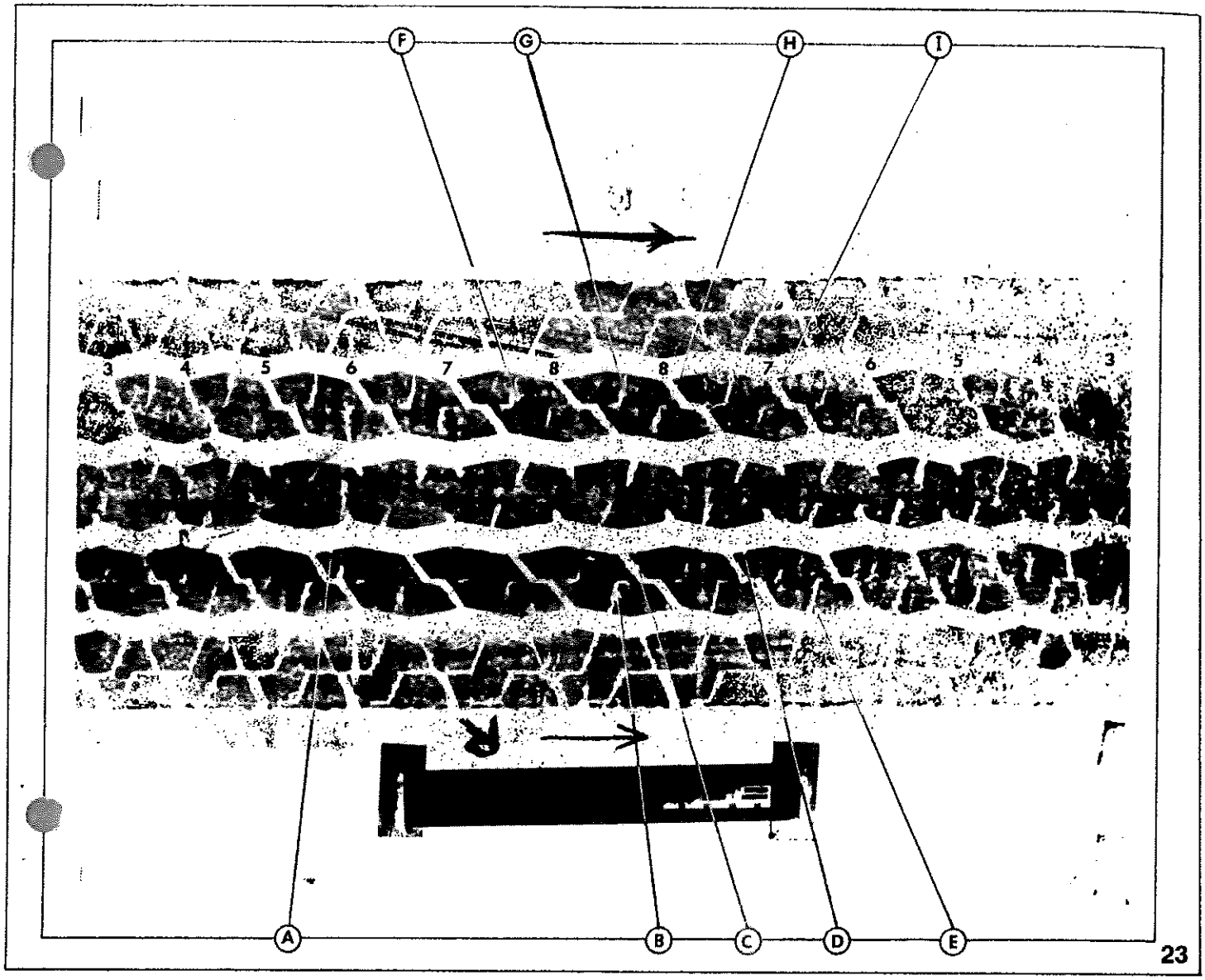
1. The Firestone & Rubber Co. — *Computerized Tire Footprint Areas Measurements*, by C.E. Prettyman.

COMPARISON OF CRIME SCENE TO TEST IMPRESSION

Using the negative of a crime scene impression, the examiner can prepare a full size photograph and Kodalith transparency. This technique works well for two dimensional and three dimensional crime scene impressions. However, in the case of a three dimensional impression, Kodaliths prepared from plaster casts work better since the cast can be photographed at the studio under controlled lighting conditions. The cast also provides very accurate dimensions for one to one enlargements. Nevertheless I have used Kodaliths prepared from all three types of impressions with good results.

In the case study used for this article a full size photograph and Kodalith transparency of the dust impression lifted with the carbon paper, was prepared for comparison to the test impression of the Super 125. The arrow on the Kodalith and photograph were lined up in the same direction as the arrow on the test impression. The transparency was then moved along the test impression to see if the noise treatment could be matched (see Figure 21).

If the dust impression lifted with carbon paper had been made by a different size tire of the same design, the noise treatment would not have fit in this case, since it would have been proportionately larger or smaller. As well, if the wear to the sipes, notches and grooves was not consistent the test impression could have been ruled out. In this case I also had a Super 125 of the same size mounted on the left rear side of the test vehicle. The left rear tire was worn to a greater extent and when its test impression was



23

lettered A to I in agreement with the dust impression lifted from the cardboard box exhibit.

Figure 25 shows the accidental characteristics lettered A to I located on the suspect tire. Small markers have been placed on the tread elements pointing to the cuts in question for quick and easy reference.

In cases involving three dimensional impressions in soil or snow I have prepared full size photographs and Kodaliths to search the suspect tire test impression. When areas of noise treatment agreement are located on the test impression I mark them for reference. In a case where no substantial wear or accidental characteristics are present there will be one or more locations which are consistent with the crime scene impression. For demonstration purposes, I have found in most cases, that if I photograph one of these areas on the test impression and prepare a full size Kodalith from the negative, and then place the test impression Kodalith over the crime scene photograph, it will better display the agreement of noise treatment. This is because the test impression has much better contrast and will produce a Kodalith which is easier to look through. This procedure creates the perspective of being inside the tire looking through the tread design and seeing how the various elements could have made the three dimensional depressions in the soil or snow.

As was already mentioned, it is very important to ensure that the ruler is in the same plane as the crime scene impression when taking the photograph. Then if care is taken in the darkroom to produce an accurate full size Kodalith it will reflect the proper dimensions for the noise treatment on the tire which made the crime scene impression. For two and three dimensional impressions recorded on a smooth surface the Kodalith made from them should produce a very accurate fit with a test impression from the same tire, since the test impression is also recorded on a flat surface. To illustrate this, I refer to a recent case which involved several large thefts of fuel from oil rig sites in the southeastern part of Saskatchewan.

my 1985 tour of Firestone Headquarters in Akron, Ohio.

Special thanks to Uniroyal Canada Ltd. for their assistance and my appreciation to employees who spent time with me on my 1983 tour of their tire manufacturing plant in Kitchener, Ontario.

Special thanks to The B.F. Goodrich Company for information and material they provided to me. Thank you to the employees of the B.F. Goodrich Company who spent time with me on my tour of their headquarters in Akron, Ohio, in 1985.

Special thanks to Mr. Pete MacDonald of Tire Forensics, Hudson, Ohio, whose sharing of information has provided me with a greater insight into the field of forensic tire impression identification.

Thank you as well to all the others who have assisted me in one way or another on my research in this area.

BIBLIOGRAPHY

1. MacDonald, Peter — *Tire Styling and Forensic Tire Footprint Identification*
2. The Firestone Tire and Rubber Co. — *Factors Affecting Passenger Tire Traction on the Wet Road* by J.D. Kelley Jr.
3. The Firestone Tire and Rubber Co. — *Computerized Tire Footprint Area Measurements* by C.E. Prettyman
4. The Firestone Tire and Rubber Co. — *From the Milk of a Tree*
5. The Firestone Tire and Rubber Co. — *For a World on Wheels*
6. *Rubber World* — "Tread Designs: Yesterday and Today" by Addis Finney
7. *Rubber World* — "All-Season Tires: A Welcome Innovation" by Addis Finney
8. Resources Development Corporation — *Tires: Performance and Construction*
9. Resources Development Corporation — *Tires: Troubleshooting and Service*
10. B.W. Given, R.B. Nehrich and J.C. Shields — *Tire Tracks and Tread Marks*
11. The Goodyear Tire and Rubber Co. — *Tire Technology* by F.J. Kovac
12. B.F. Goodrich Tire Co. — *Research and Developments T/A High Tech Radials*
13. *Popular Mechanics* — "Technology Comes to Tire Design" by Wayne W. Williams
14. Uniroyal Canada Ltd. — *Rubber and Tire History. Tire Manufacturing Process. Tire Construction. The Radial Tire. Tire Terminology. Tire Wear.*
15. Grogan and Watson — *Tyres and Crime*
16. Grogan, R.J. — *Tyre Marks as Evidence*, Dunlop Ltd.
17. Rubber Manufacturers Association — *Care and Service of Automobile and Truck Tires*
18. The General Tire & Rubber Co. —*Information on Tire Manufacturing*

The Lighter Side — Kiss Wasn't Assault

Reprinted from the Police Review, May 4, 1984 edition, London, England

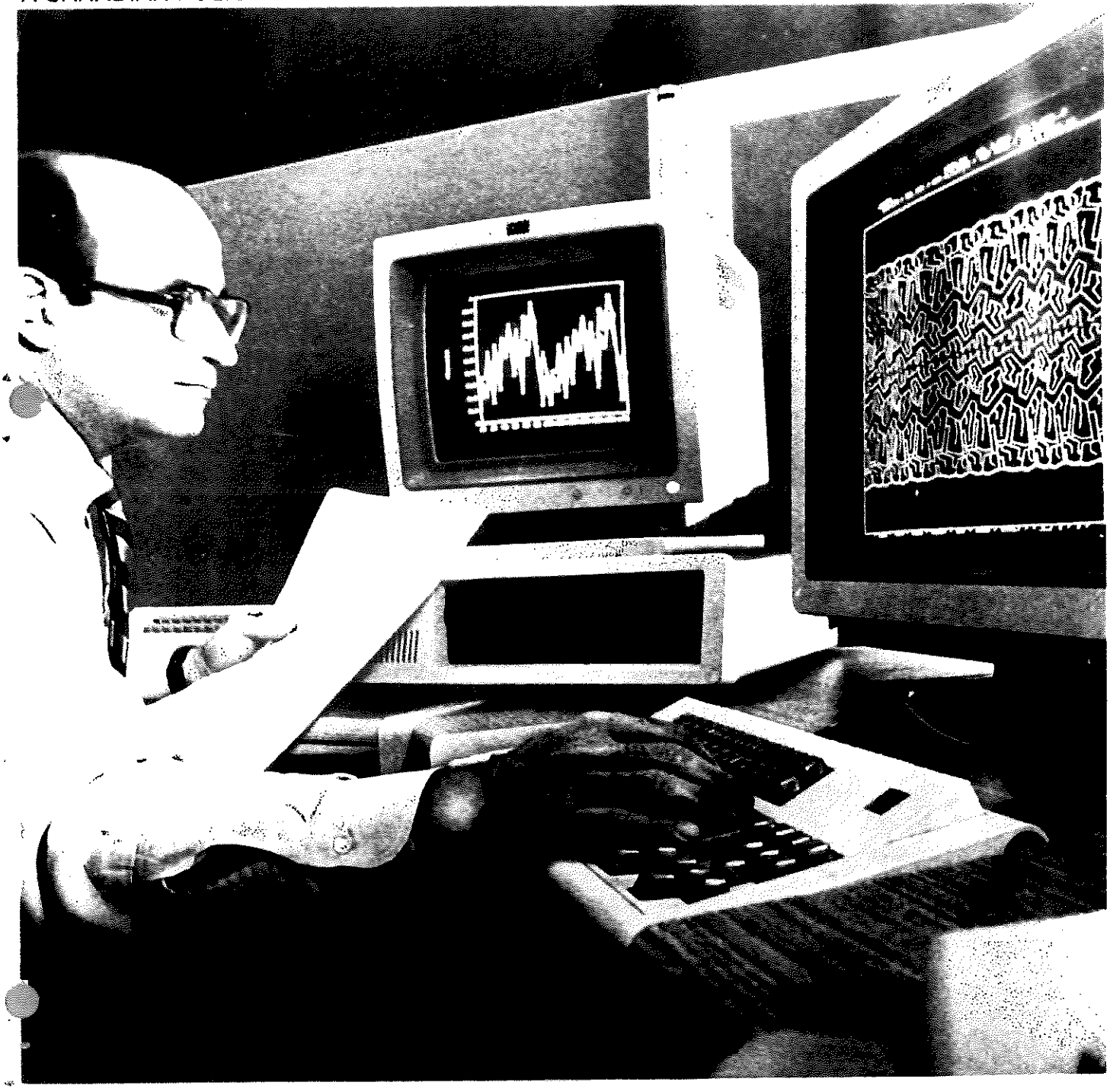
A baker who kissed a policeman on the lips was held to be not guilty of an assault on police. The sheriff for the Highlands and Islands was giving judgement in a case in which police alleged that the baker, after being given help with his broken-down car, became abusive, and was put into the back seat of the police car.

PC Derek Mitchell said that the baker continued to abuse him, then started blowing kisses at him, and finally kissed him on the lips.

The *Glasgow Herald* quoted from the sheriff judgement: "It is impossible to say from the evidence whether this kiss was an act of retaliation, a spontaneous gesture of affection, or an irresistible sexual impulse. . . . Mercifully, the nature, extent, and vigour of the kiss planted on PC Mitchell's lips was not explored in evidence. I shall regard it as an unusual but justifiable act of retaliation which I fervently hope does not become a general practice."

FRONT AND REAR COVERS: This issue is devoted to the science of tire impression identification (see page 1).

A CANADIAN POLICE SERVICE VOL. 49, No. 1, 1987



From: [redacted]
To: [redacted]
Subject: Response to: PCAST Call for Additional References regarding the PCAST Report on Forensic Science 2016 (UNCLASSIFIED)
Date: Wednesday, December 14, 2016 6:22:40 PM
Attachments: PSAC-FR_PCAST response_20161214.pdf

CLASSIFICATION: UNCLASSIFIED

Dear Mr. Eric Lander & PCAST colleagues,

The Organization for Scientific Area Committees, Friction Ridge Subcommittee respectfully submits the attached response to the PCAST "Call for Additional References" regarding the 2016 PCAST Report on Forensic Science.

We appreciate the attention the PCAST has given to this matter and are available if we may be of any further assistance.

Sincerely,

Henry Swofford
Vice-Chair, Friction Ridge Subcommittee
Organization for Scientific Area Committees

CLASSIFICATION: UNCLASSIFIED

Organization of Scientific Area Committees
Friction Ridge Subcommittee

Response to Call for Additional References Regarding:

President's Council of Advisors on Science and Technology
REPORT TO THE PRESIDENT
Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods
14 December 2016

The Organization of Scientific Area Committees (OSAC) Friction Ridge Subcommittee (FRS) is thankful for the attention of the President, and other relevant members of the Executive Office, including the President's Council of Advisors on Science and Technology (PCAST), to ensure forensic science methods are adequately resourced, properly evaluated, and appropriately applied in practice to safeguard the validity of forensic evidence used in the Nation's legal system. The OSAC, a body of over 500 forensic science practitioners and other experts who are drawn from local, state, and federal agencies; academia; and industry, share this concern and demonstrate commitment by voluntarily serving on various subcommittees with the overarching intent to develop and promulgate forensic science consensus documentary standards and guidelines, and to ensure a sufficient scientific basis exists for each discipline. By nature of its mission, the OSAC FRS acknowledges, and agrees with the PCAST, that there is a need to establish a more formalized research agenda, develop standardized practices, and promote foundational research. Overall, the OSAC FRS considers the PCAST report to be well written and clear with respect to their evaluation criteria and strategic way forward. The aspects of the report the OSAC FRS believes could be further clarified and elaborated upon are:

(1) The OSAC FRS agrees with the PCAST that there is a need for additional research to build upon an established body of knowledge; however, we disagree that prior research efforts should be disregarded or discounted in their entirety.

(2) The PCAST states black box studies are the *only* means of establishing foundational validity for subjective feature-based methods and "[i]n the absence of such studies, a subjective feature-comparison method cannot be considered scientifically valid". (p. 66). While the OSAC FRS agrees with the need for black box studies to evaluate the overall validity of a particular method, the OSAC FRS is concerned this view could unintentionally stifle future research agendas aimed at dissecting the components of the black box in order to transition it from a subjective method to an objective method. If the PCAST maintains such emphasis on black box testing as the *only* means of establishing validity, the forensic science community could be inundated with predominantly black box testing and potentially detract from progress on understanding and refining other foundational aspects of the method, such as those previously outlined by the OSAC FRS, in an effort to identify ways in which to emphasize objective methods over subjective methods (see https://www.nist.gov/topics/forensic-science/osac-research-development-needs). Given the existing funding limitations, this will be especially problematic and the OSAC FRS is concerned other foundational research will thus be left incomplete.

(3) The OSAC FRS notes that the PCAST appears to discount or otherwise disregard the role of "experience" and "judgment" in subjective feature-comparison methods. While the OSAC FRS

does value empirical testing as hierarchically greater than experience and judgment, they do play an important role and should not be disregarded in their entirety. The disregard for experience and judgment is reminiscent of what was initially proposed by the Evidence Based Medicine (EBM) movement in the early 1990's. Cohen et al. (2004) in the context of medicine, cautions that experience and judgment remain important elements, especially in situations in which the circumstances of a particular case is underrepresented by empirical tests. (see Cohen et al., "A Categorization and Analysis of the Criticisms of Evidence Based Medicine". *International Journal of Medical Informatics,* 73 (2004) 35-43). Similar principles may be applied to the forensic sciences. Accordingly, the emphasis should not be on blindly banning knowledge gained and assessments made on the basis of experience and judgment; rather, the emphasis should be on clearly distinguishing the source of such knowledge and transparently reporting its basis and associated limitations.

(4) The PCAST considers the error rate for latent fingerprint analysis to be "substantial" with estimates that an error may be expected to occur up to 1 in every 306 cases (based on the FBI/Noblis study) and 1 in every 18 cases (based on the Miami Dade study). Further, the PCAST states "the actual false positive rate in casework may be higher" (p. 101). The OSAC FRS has some concerns with the approach used by the PCAST to arrive at those estimates.

a. The PCAST based their quoted estimates on only a subset of the examination methodology. It is common practice within the latent fingerprint community to ensure conclusions have been verified by a separate examiner prior to a conclusion being released. While the OSAC FRS recognizes that many laboratories may not perform "blind" verifications, the error rates quoted by the PCAST did not consider *any* verification being performed. Accordingly, the error rates quoted by the PCAST do not necessarily reflect actual casework methodology. Both the FBI/Noblis study and the Miami Dade study demonstrate that false positive errors reported by one examiner were rarely reproduced by a second examiner. Taking this into consideration, in practice, the error rate is expected to be lower, perhaps to a substantial degree, than those values reflected by the PCAST.

b. The PCAST stated "because examiners were aware they were being tested, the actual false positive rate in casework may be higher" (p. 101). Based on that statement, it appears the PCAST was referring to the Observer "Hawthorne" Effect. The Hawthorne Effect suggests that test subjects may modify or improve an aspect of their behavior in response to their awareness of being observed. The implication of the PCAST is that the error rates observed in the study may represent a lower error rate than what may be expected in casework. While the OSAC FRS recognizes the Hawthorne Effect could be applicable in this situation, the OSAC FRS also notes the research was conducted in an anonymous fashion (as appropriately required by Institutional Review Boards) and participants may paradoxically behave less accurately when they know their identity is concealed and there are no downstream consequences to an incorrect response. Lelkes et al. (2012) observed this phenomenon stating that total anonymity "consistently reduced reporting accuracy and increased survey satisficing [and] complete anonymity may compromise measurement accuracy rather than improve it." (see Lelkes et al., "Complete Anonymity Compromises the Accuracy of Self-Reports". *Journal of Experimental Social Psychology*, 48 (2012) 1291-1299). Although the OSAC FRS does not have empirical evidence to substantiate what the *actual* rate of error is in practice, the OSAC FRS believes strongly that it is not on the level of magnitude reported by the PCAST. If this were true, considering the prevalence of friction ridge evidence examined

on an annual basis around the country, the criminal justice system would be inundated with non-corroborative evidence which would draw considerable attention to the issue.

c. The PCAST considers the quoted error rates for the latent fingerprint discipline as a combination of both human/technical failures and coincidental matches and recommends those error rates be reported to the courts. While the OSAC FRS supports the suggestion to provide error rates to the courts, those rates should be accurately calculated, relevant to the circumstances of the individual case, and appropriately articulated. The error rates quoted by the PCAST are generalized across a sample set of latent fingerprints in which their qualities represent the least favorable conditions that may be observed in casework. Accordingly, while on *average*, the quality of samples utilized in those studies may be consistent with the *average* quality of "difficult" or "complex" samples examined in casework, the error rate should not be generalized as a single rate of error for all latent fingerprint casework; rather, the error rate should be relevant to the quality of the fingerprint *in the case at hand*, as noted by the PCAST with the statement, "[t]he false positive rate for latent fingerprint analysis may depend on the quality of the latent print." (p. 50). The OSAC FRS agrees with the PCAST that error rates should be conditioned upon the quality of the fingerprint sample and encourages this research to be carried out. In the interim, the OSAC FRS believes it is appropriate to inform the fact-finder that the error rate *in the case at hand* may actually be lower than those observed in the black box studies considered by the PCAST. The OSAC FRS believes this is appropriate because the error rates quoted by the PCAST were calculated on the basis of an incomplete methodology (see sub-section 4a), under conditions which do not reflect actual casework (see sub-section 4b), and are not conditioned on the quality of the fingerprint sample *in the case at hand*. Taking these points into consideration, if the friction ridge community were to report error rates quoted by the PCAST without providing appropriate context, the friction ridge community could unduly bias the fact-finder to either undervalue the "true" value of the evidence (in the case of very high quality evidence) or overvalue the "true" value of the evidence (in the case of very low quality evidence).

d. The PCAST relies heavily on the "Miami-Dade" black box study as a means of estimating the error rate for the latent fingerprint discipline. The OSAC FRS notes that the PCAST failed to detect the calculation error in the false positive rate reported by Miami Dade. The false positive rate is calculated as the number of false positive responses divided by the number of opportunities to make a false positive response (conditions in which non-mated samples were presented to the study participant). The Miami-Dade study differed from the FBI/Noblis black box study in that the FBI/Noblis study consisted of a single latent impression compared to a single reference impression (hence a false positive could only occur in a non-mated trial). The Miami-Dade study, on the other hand, provided participants with *multiple* reference impressions for each trial; thus, even for the trials which contained a mated source, there were also non-mated sources which could have resulted in a false positive response. Indeed, of the 42 false positive results, 39 of them were made to an incorrect reference print during a mated source trial. Accordingly, the accurate calculation for the false positive rate in the Miami-Dade study is 42 false positive responses divided by 3,687 trials in which a false positive response could have occurred and in which a conclusive response was rendered (1,398 non-mated source trials and 3,138 mated source trials with non-mated source reference prints minus 849 inconclusive decisions among both mated and non-mated sets). Rather than a false positive rate of 4.2% (42/995), as stated by the authors and quoted

by the PCAST, the actual false positive rate is 1.1% (42/3,687). The upper bound of the 95% confidence interval then becomes 1.5% (not 5.4% as originally calculated by the PCAST). Further, if the 35 false positive responses believed to be due to clerical errors were removed, the observed false positive rate is 0.19% (7/3,687). The upper bound of the 95% confidence interval then becomes 0.39%.

(5) The PCAST states "[s]ubjective methods can evolve into or be replaced by objective methods." (p. 47). The OSAC FRS recognizes the integration of objective methods to measure similarity compared against a pre-defined "matching" criteria will certainly be a step in the right direction; however, the OSAC FRS believes it is a mistake to expect that objective methods will *fully* replace the subjectivity of the human examiner. The human examiner will continue to serve as a critical, albeit subjective, element of the broader methodology. Rather than entire substitution, the human examiner and the measurement instrument will need to work complementary to one another. This is how science of all sorts is practiced.

In closing, the OSAC FRS appreciates the attention and commitment to improving forensic science demonstrated by the PCAST, the President, and other members of the Executive Office. The OSAC FRS believes the forensic sciences and, latent fingerprint analysis in particular, are on the right path forward to build upon an existing foundation of knowledge, improved standards and guidelines, and strategies to transition elements of the methodology which rely on subjective judgment into objective measurements. The OSAC FRS looks forward to a joint commitment to this effort by the general scientific community.

From: Robert Dorion
To: [redacted]
Subject: Reply to the "Forensic Science in the Criminal Courts: Ensuring Scientific Validity Of Feature-Comparison Methods" report
Date: Wednesday, December 14, 2016 7:56:48 PM
Attachments: Dorion in response to PCAST invitation to reply 20161214.pdf

---

Eric Lander
Co-Chair, PCAST

As requested, herein is in response to PCAST invitation to reply to the "*Forensic Science in the Criminal Courts: Ensuring Scientific Validity Of Feature-Comparison Methods*" report.

Respectfully,

Dr Robert B.J. Dorion

December 13, 2016

Dear PCAST,

You have invited additional commentary on your Sept 2016 report, "Forensic Science in the Criminal Courts: Ensuring Scientific Validity Of Feature-Comparison Methods." Our comments will focus specifically on DNA.

You ask:

> Please identify any relevant scientific reports that (i) have been published in the scientific literature, (ii) were not mentioned in the PCAST report; and (iii) describe appropriately designed, research studies that provide empirical evidence establishing the foundational validity and estimating the accuracy of any of the following forensic feature-comparison methods, as they are currently practiced:
>
> DNA analysis of mixed samples with three or more contributors, in which the contributor in question represents less than 20% of the sample.
>
> Please indicate how the scientific reports establish foundational validity and estimate the accuracy of the relevant method.

Our group (Professor Keith Inman, Dr. Kirk Lohmueller, Dr. Norah Rudin) has been working specifically on this problem over the last several years. The work has been funded by NIJ grant #2013-DN-BX-K029. We will be able to provide our final report when it is completed at the end of this month. In the meantime, we attach a draft of the abstract and executive summary. We also attach our published papers to date that address these issues. Initial results and conclusions that specifically address your requests will be discussed in the report. Manuscripts that will expand on the work will be submitted to peer-reviewed journals over the coming year. Importantly, the curated data set that has been generated, which comprises 800 complex samples of up to 4 contributors, and over a wide variety of template amounts and contributor ratios, will be made publicly available.

We have several further questions and comments regarding the final PCAST report.

In the report you cite a publication that mentions validating mixtures of which the minor contributor is 20% or more of the sample. Could you please specify the publication on which you are relying for this statement?

On page 75 of the PCAST report, you offer a definition of a complex mixture as follows:

> *DNA analysis of complex mixtures—defined as mixtures with more than two contributors—is inherently difficult and even more for small amounts of DNA.*

The citation you reference is the SWGDAM guidelines (2010).

We suggest a more inclusive definition of a complex *sample* (not merely a *mixture*) as one that exhibits one or more of the following characteristics: (Butler 2015)

- A low template sample containing DNA from one, or more than one, individual in which dropout may have occurred
- A mixture of two or more individuals
- A sample that suffers from degradation
- A sample that suffers from PCR inhibition

Each of these circumstances conspires to introduce additional ambiguity that further complicates interpretation of the sample. Essentially any sample for which the full and complete genotype of individual donors cannot be determined with certainty should be considered a complex sample. Probabilistic genotyping approaches are designed to model these types of ambiguous profiles. No binary approach, RMP, CPI or otherwise, can adequately account for the ambiguity inherent to these samples. Omitting, for example, a low template single source sample with the possibility of dropout from your definition implies, incorrectly, that a CPI is an appropriate method of providing a weight of evidence for such a sample.

We were surprised and concerned to note the substantive addition that occurred between the draft and final PCAST report. Specifically, the conclusion that the CPI should be summarily dispensed with is followed by an edict that allows it to be used following the procedure outlined in Bieber et al. 2016 (Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculation using the combined probability of inclusion). Even more surprising, PCAST endorses this paper without, by their own admission, having adequately reviewed it. This does not seem in keeping with the rest of the well-crafted report. We have read the Bieber et al. paper and find it to be at odds with the basic scientific premise that PCAST espouses, e.g., that procedures should be foundationally valid, and based on ground truth samples. The Bieber et al. paper lacks any foundation in experimentation based on ground truth samples. Instead, it just continues the unfortunate legacy of binary-type approaches (CPI, CPE, RMNE, 2P etc.), none of which have ever been subjected to large-scale, rigorous validation. Although many forensic DNA practitioners agree, at least in theory, that probabilistic genotyping must replace binary methods, the agreement is not universal. Resistance remains, and there are those who will look for any endorsement to support the continued use of historical approaches. Our fear is that this unfortunate last minute addition to the PCAST report will allow the CPI and similar approaches to be used for the foreseeable future, as they now have the imprimatur of PCAST.

Keith Inman M.Crim

Norah Rudin, Ph.D.

Kirk Lohmueller, Ph.D.

References

Haned, H., Gill, P., Lohmueller, K., Inman, K., Rudin, N.; Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations, *Science and Justice* 56 (2016) 104–108

Marsden, C.D., Rudin, N., Inman, K., Lohmueller, K.E.; An assessment of the information content of likelihood ratios derived from complex mixtures; *Forensic Science International: Genetics* 22 (2016) 64–72

Bieber et al., Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion, *BMC Genetics* (2016) 17:125

Butler, J.M. "The future of forensic DNA analysis." *Philosophical Transactions of the Royal Society B,* 370: 20140252 (2015)

Scientific Working Group on DNA Analysis Methods, SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories, 2010 Available from: http://www.fbi.gov/about-us/lab/codis/swgdam.pdf.

Science and Justice 56 (2016) 104–108

Contents lists available at ScienceDirect

Science and Justice

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

Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations

Hinda Haned [a,*], Peter Gill [b,c], Kirk Lohmueller [d], Keith Inman [e], Norah Rudin [f]

[a] Netherlands Forensic Institute, Department of Human Biological traces, The Hague, The Netherlands
[b] Norwegian institute of Public Health, Oslo, Norway
[c] Department of Forensic Medicine, University of Oslo, Norway
[d] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, 621 Charles E. Young Drive South, Los Angeles, CA 90095-1606, United States
[e] Department of Criminal Justice Administration, California State University, East Bay, 4069 Meiklejohn Hall, 25800 Carlos Bee Boulevard, Hayward, CA 94542, United State
[f] 650 Castro Street, Suite 120-404, Mountain View, CA 94041, United States

ARTICLE INFO

Article history:
Received 18 June 2015
Received in revised form 26 November 2015
Accepted 27 November 2015

Keywords:
Validation
Likelihood ratio
Probabilistic
Genotyping
Strength of evidence
LRmix

ABSTRACT

A number of new computer programs have recently been developed to facilitate the interpretation and statistical weighting of complex DNA profiles in forensic casework. Acceptance of such software in the user community, and subsequent acceptance by the court, relies heavily upon their validation. To date, few guidelines exist that describe the appropriate and sufficient validation of such software used in forensic DNA casework. In this paper, we discuss general principles of software validation and how they could be applied to the interpretation software now being introduced into the forensic community. Importantly, we clarify the relationship between a statistical model and its implementation via software. We use the LRmix program to provide specific examples of how these principles can be implemented.

1. Background and scope

A number of new computer programs have recently been developed to facilitate the interpretation and statistical weighting of complex DNA profiles in forensic casework (for a review, see [1]). Complex profiles may encompass a multitude of confounding factors resulting from DNA profiling of a low quantity and/or low quality biological sample. The resulting profile may contain multiple contributors, may lack information from the true contributors (allelic drop-out), may include extraneous information unrelated to the crime-sample information (allelic drop-in), and may suffer from degradation or inhibition [2].

It is now accepted throughout the world-wide forensic DNA community that a likelihood ratio (LR) approach is required to reliably interpret these types of profiles [3]. Accordingly, recent years have seen a proliferation of probabilistic models, implemented via software, offered to the community as solutions to this problem. Although these probabilistic models rely on different assumptions, and make use of different types of information, they all enable the evaluation of evidence within a LR framework. While these software programs have proven generally useful to facilitate the interpretation of complex DNA profiles, [4–7], no generally accepted guidelines exist to establish their validity for use in forensic casework. Model validation for use in forensic casework is not straightforward because the true weight of the DNA evidence cannot be determined; indeed, the generated LR always depends on the model's assumptions, no 'gold standard' exists in the form of a true likelihood ratio that can serve as a comparison [8,9].

In this paper, we offer a set of definitions and examples that aim to provide guidance in validating software for casework use. We first introduce some general definitions of model and software validation taken from existing fields. We then propose a set of considerations for validating software for forensic use. We illustrate the application with the LRmix program [10], which has been validated for casework use and introduced into a courtroom setting.

2. Definition of validation

Forensic science is not the first discipline to face the challenges of model and software validation. Consequently, it is possible to learn from the experience of scientists working in different fields. We follow Rykiel [11] in his definition of model validation (originally applied to the field of ecological science). This paper is highly cited and is effectively regarded as a 'standard reference'. We regard model validation as a

* Corresponding author at: Netherlands Forensic Institute, Netherlands.

http://dx.doi.org/10.1016/j.scijus.2015.11.007
1355-0306/

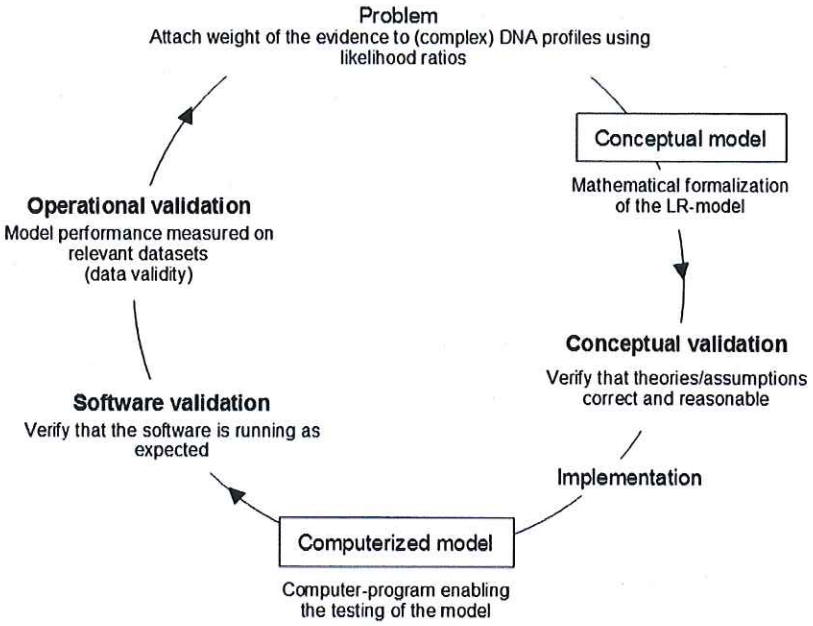


Fig. 1. Simplified representation of the model development and validation process. The diagram shows the different stages of conceptual, operational and software validation (modified from [11]).

process that results in an explicit statement about the behavior of the model (and subsequently the software). In the case of an interpretation model, such a statement would be: "The implementation of Model X in Software Y is valid for application in forensic casework subject to limitations described in the operational validation document".

Model and software validation are inherently entangled, as software implementation is always needed to implement and use a model (see Fig. 1). However, the two concepts can be related in a simple way; the software is merely a vector for the model. As illustrated in Fig. 1, validated software can actually rely on an invalid model, for example, if the underlying theory or mathematics are shown to be flawed. The goal is to implement a valid model, but it is important to realize that correct implementation of the mathematics of a model by a piece of software provides no information about the validity of the model itself; conversely, demonstration of correct implementation is a critical part of validation.

2.1. Model validation

Model validation ensures that the model has been extensively checked to be sound and fit for purpose. This can be achieved through two steps: conceptual validation and operational validation [11].

2.1.1. Conceptual validation

Conceptual validation verifies that the mathematical formalization of the model, as well as its underlying assumptions, is fundamentally correct. Publication of the theory of the model in peer-reviewed scientific journals allows an opportunity for the underlying theory to be independently assessed, articulates the underlying assumptions, and, most importantly, documents the scientific support for the model structure. For this step to be successful, the model theory must be thoroughly explained. Publication, while necessary, is not sufficient; an editorial decision to publish a paper does not constitute fundamental proof of the scientific validity or usefulness of the contents.

The advent of electronic publication removes space restrictions and allows for the possibility of publishing online supplementary material, and gives modellers the opportunity to expand on their methods. The underlying data on which the conclusions are based can and should be published as supplementary material so that independent researchers can inspect it and use it to independently verify the results obtained. For open-source software, the computer code can also be published as supplementary material, or as a link provided to the location of the code [12]. The code can then be studied by independent researchers, facilitating an understanding of the model, an important component of conceptual validation. The implementation of the model can also then be independently assessed by interested parties.

The most straightforward way to demonstrate conceptual validity is for the model developer to embrace a transparent approach, which allows for true independent review and verification. A transparent approach requires all of the model assumptions to be described, and accessible to anyone who wishes to independently re-implement the model. This approach is demonstrated by [7,8]. This is diametrically opposed to a black-box approach in which only partial explanations are provided, denying an independent researcher the ability to scrutinize the details and re-implement the model if desired [3,13].

2.1.2. Operational validation

We follow [14] and define operational validation as the procedure that determines whether "the model's output behavior has the accuracy required for the model's intended purpose over the domain of the model's intended applicability". Operational validation is usually verified using a "computerized model". In other words, unless a computer implementation of the model is available that can run a profile and yield an output, the operational validity of the model cannot be tested (Fig. 1). Operational validity is tested via user-defined criteria that can be either accepted or rejected. These can be determined for LR-based models. For example, the following properties can readily be tested:

- Comparison to a standard basic model that operates with minimal assumptions so that the effectiveness of models that take into account additional parameters may be measured objectively. Gill and Haned [15] defined the requirements for such model, which allows the evaluation of complex DNA profiles without using all available information.
- The LR of a set of propositions for any profile is lower or equal to the inverse match probability of the profile questioned under the numerator hypothesis [9].
- The LR obtained for a given profile decreases with increasing

ambiguity and decreasing information content [9]. Specifically, any deviation from a one-to-one correspondence of the suspected contributor profile and the evidence profile, as well as any loss of information from the evidence profile itself, should reduce the LR.

- The LR can be compared to a benchmark LR value. A benchmark LR can be calculated when most parameters of the model can be estimated from known profiles (see below). The reasons for any differences between the observed and expected output can be investigated and the model can subsequently be modified to yield the expected output.

2.1.3. Defining benchmarks for LR-based models

Benchmark likelihood ratios can be calculated for certain models for which parameters can be estimated directly from samples with known input. The LRs obtained with parameters estimated from such samples, and the LRs calculated with the estimates for another test dataset should converge [2]. The quality and range of the data used for operational validation are critical [11,14]. We follow Sargent [14] and define data validity as "ensuring that the data necessary for model building, model evaluation and testing, and conducting the model experiments to solve the problem are adequate and correct". Typically, experimental data sets for which the true composition of the samples is known are used (see for example [7,16]). Test samples chosen should represent the spectrum of situations encountered in real-world casework. Profiles representing extreme situations should be included, even if these profiles ultimately might not be interpreted in casework. The idea is to determine not only when the system works as expected, but also when it may fail. Specifically, it is important to investigate the boundaries of the model within its domain of application. Common characteristics of forensic casework samples that can increase their complexity include multiple contributors, low quantity (provoking possible drop-out) and low quality (e.g., degradation, inhibition, contamination). All of these factors increase ambiguity and reduce information content. Both the limitation of the model and the limitations of the evidence must be tested. For example, validation may determine that, past a certain number of contributors, the information content of the profile is simply too limited to reliably distinguish a true contributor from a non-contributor who shares some of the detected alleles by chance. Therefore, based on an operational validation of the model, as implemented by software, it might be relevant to impose a limitation on attempting to interpret casework samples that exceed some defined number of contributors to a mixture. Simulated data can prove helpful in exploring model limitations; however, they cannot substitute for experimental data [13]. Any parameters modeled using simulated data must always be tested on profiles generated from physical samples, and the model refined based on the outcome. The most robust models are those tested with the widest range of data [14]. This is well illustrated by Nordstrom: "The greatest weaknesses of any model computation are the quality of the input data and the adequacy of the assumptions (implicit and explicit); remember GIGO ('garbage in, garbage out')."

2.2. Software validation

Model and software validation usually are carried out simultaneously, as it is the computerized version of the model that enables the model validation exercise (Fig. 1). We define software validation as ensuring that the programmed algorithms follow the mathematical concepts defined in the model. We suggest the following main steps for software validation:

1. Define the statistical specifications of the software: This is an outline of the theory behind the model to be implemented in the computerized version of the model. This document compiles the information that is typically available in peer-reviewed papers describing the model and software implementation.
2. Carry out analytical verification: For example, analytical calculations of likelihood ratios using simple cases (e.g., single-source and two-person mixtures) can be derived and compared to the software output. Depending on the complexity of the model, analytical verification may or may not be possible. This has been termed the "complexity paradox" by [13]; the more complex a model is, the more difficult it is to verify the different blocks of the model. In such a case, the software output can be compared to output from alternative software that implements a similar model.
3. Compare to parallel implementations: Comparisons to alternative software, either relying on a similar or a different probabilistic model, can be useful to verify software behavior. Such comparisons rely on the 'convergence principle' described by Gill and Haned [15], as well as Steele and Balding [1]. Numerical differences between software, corresponding to one unit on the log_{10} scale are negligible [1].
4. Verification of the code itself through visual inspection and recoding. This is most easily achievable through open-source software.

3. Validation in practice

In Box 1 we illustrate an example of validation in practice using the LRmix program, which is freely available in the Forensim R package [17]. While the general approach to validation is applicable to all systems, the specifics will vary depending on the model and the variables that are included. Validation of the model itself will concentrate on the variables that add information content. Questions important to

Box 1
Validation steps for the LRmix program for use in forensic casework.

Step 1. Conceptual validation
- Model theory and assumptions were explained and justified in a "statistical specifications" report,
- Model theory was formalized and published in peer-reviewed journals [18,19].

Step 2. Operational validation
- Model output was compared to expert opinion on 20 cases,
- Model output was compared to the following programs: Lab Retriever [20], LikeLTD [21], FST [16], GRAPE [22],
- Performance tests using 211 controlled mixtures (of one up to five contributors) and 621 (overall-loci) likelihood ratios were compared to expected trends based on gold standard conditions where parameters were known.

Step 3. Software validation
- Software output was evaluated analytically, using the Xcas algebra software
- Model output was evaluated on 77 controlled NGM mixtures, and >1000 LRs were computed and compared to expected trends using known parameters
- LRmix output was evaluated against analytical formulae derived for simple examples
- LRmix output was evaluated against an independent re-implementation of the model (in the Java language), using 77 controlled NGM mixtures, and >1000 LRs were computed and compared

Validation statement: "Over the 1095 LR calculations were submitted to comparisons of LRmix and other software, for all tested samples, the same conclusions were obtained. We therefore concluded that LRmix is validated for use in casework, within the limitations described in the operational validation document."

this process include: how does changing the values for these variables change the LR; at what point does it make a significant difference; what is the effect of using extreme values; and when does the model/software behave in an unexpected way? Box 1 outlines the different steps carried out to validate LRmix for use in forensic casework. These steps are given as an illustration on how the validation procedure might be carried out in casework.

4. Discussion

Although no guidelines yet exist on the best methods to validate forensic DNA interpretation software for casework use, we can draw on the collective wisdom of other disciplines to guide our inquiry. We can also comment on published validation efforts of software tools that have been offered for the interpretation of forensic DNA profiles.

4.1. Does the validation answer the relevant scientific question(s)?

The DNA Commission of the International Society for Forensic Genetics recommendations [3] offers some suggestions for best practice, including a transparent approach that readily allows for independent verification. The Scientific Working Group in DNA Analysis Methods (SWGDAM) has convened a probabilistic genotyping subcommittee that produced a document of concrete guidelines in 2015 [23].

While software validation might appear to be straightforward, model validation may lead to epistemological questions about the true meaning of a validated model. Here, we argue that model validation is possible, given a particular context of use, within a specific framework of limitations, and for an explicit implementation. One of the limitations to consider is the complexity of the model. The more complex the model, the greater the number of assumptions that are required. Increasing the number of variables incorporated into such a model also increases the chance of creating dependencies. Such models require a validation protocol that specifically addresses the additional interactions, and care must be taken to clearly define the variables. We caution that complex models may at some point begin to produce unrealistic results, and hence become counter-productive. More generally, the validation criteria should be explicit to the end users, and a determination made as to whether these criteria are fit for purpose. Within a coherent quality framework, the criteria may be improved over time. As an example, the steps used to validate LRmix are provided to users (Box 1).

4.2. Software and model comparisons

Comparison of the model to be validated to other models is an important part of validation. However, this can be difficult in practice due to differences between the models themselves. Therefore, attempting to set parameters to exactly the same values for each system to perform a fair comparison is not always feasible, as different models rely on different variables and parameters. Typically, imposing values, or including variables, that optimize one system, especially at the expense of other variables important to another system, may produce a misleading comparison.

Such comparisons require careful thought about which variables are important to a model and which, for the sake of the most informative comparison, must be implemented as specified by the model. As a general guideline, external factors, such as the sample population, allele frequencies and, of course, the specific hypotheses compared, can and should be kept constant. However, variables that are used differently between the models, such as analytical threshold, peak heights, drop-out model, and even population sub-structure models, must be implemented as originally intended by the architects of the model and software.

4.3. What are the validation responsibilities of the software developer and of the end user?

The extent and type of user validation depends on the credibility of the model and the software implementing it, and this is closely related to the available information (scientific papers, tutorials, websites, seminars and workshops). Any program that uses laboratory-specific data to calibrate input variables requires at least some work on the part of the end-user. Internal validation can also be understood as an important exercise that helps the end-user familiarize himself with the software. An important aspect of this exercise is to identify and understand results that may appear counter-intuitive based on previous experience of the analyst. Assuming the model and software implementation are valid, logical explanations for these results can be found in the details of the calculations. Working through these examples can contribute greatly to the understanding of the scientist [24].

4.4. Is validated software always valid?

Models and software are dynamic; they evolve and improve over time [13]. For example, software validated for STR kits may not be used for SNP markers without an entirely separate validation exercise. This is particularly true for those models relying on empirical data, as such models rely heavily on calibration for their deployment in casework. Casework implementation might also give rise to situations that were not tested during the validation phase; these untested conditions should be submitted to the appropriate validation tests. Similarly, as software are developed or evolve, they can be tested and validated against a repository of simulated and case examples specifically prepared for such a purpose. This ensures that changes to the software are tracked and thoroughly checked.

4.5. The case for transparent software

Adopting a transparent approach is desirable when developing software for use in forensic casework [3,13]. This could be achieved in several ways. An informative discussion on the matter of validation is provided by Nordstrom [13]. Although the author uses examples from geochemistry, we believe the concepts and discussions are also relevant to forensic science. We start with this statement: "Any computer code that is used for regulatory or legal purposes must be transparent" [13]. Freely available, open-source software is a straightforward way to achieve transparency in science, as results obtained with a given software can be verified and reproduced independently [12]. Commercial software can achieve sufficient transparency if the developers choose to provide adequate information about the validation and the performance of the models. It was previously suggested that open-source software can be used as a vehicle to compare the performance of various software, including commercial software [15].

Concerns about the reliability and reproducibility of software used in scientific computing have grown over the last few years [12,25]. There is a strong movement for researchers to make the source code used for analyses freely available to the community at the time of publication. Easily accessible source code implementing a statistical method will allow scientists to perform all aspects of software validation. Availability of code will allow for operational validation as users can apply the method to known samples. Furthermore, independent researchers can visually examine the code to assess the specific implementation of the model. Finally, such transparency will promote standardisation and will facilitate improvements and extensions to existing software which will be a further benefit to the community.

5. Concluding remarks

In 1984, McCarl [24] stated "There is not, and never will be, a totally objective and accepted approach to model validation." More than

30 years after this statement was made, no generally-accepted method to test the validity of models and algorithms exists, especially in the field of probabilistic genotyping. We hope that the examples and definitions given in this paper will assist both software developers, as well as the end-users in the forensic community, to create and validate interpretation software. It is our hope that the availability of those tools will, in turn, facilitate the introduction LR-based methods for the interpretation of (complex) DNA profiles.

Acknowledgments

PG has received funding support from the European Union Seventh Framework Programme (FP7/2007-2013), Euroforgen-NOE, under grant agreement no. 285487.

References

[1] C.D. Steele, D.J. Balding, Statistical evaluation of forensic DNA profile evidence, Annu. Rev. Stat. Appl. 1 (2014) 361–384.
[2] P. Gill, H. Haned, O. Bleka, B. Hansson, G. Dorum, T. Egeland, Genotyping and interpretation of STR-DNA: low-template, mixtures and database matches—twenty years of research and development, Forensic Sci. Int. Genet. (2015).
[3] P. Gill, L. Gusmão, H. Haned, W. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P. Schneider, B. Weir, DNA commission of the international society of forensic genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Sci. Int. Genet. 6 (2012) 679–688.
[4] K.E. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (2013) S243–S249.
[5] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Alemañ, M. Aler, F. Álvarez, C. Baeza-Richer, A. Dominguez, C. Doutremepuich, M. Farfán, M. Fenger-Grøn, J. García-Ganivet, E. González-Moya, L. Hombreiro, M. Lareu, B. Martínez-Jarreta, S. Merigioli, P.M. del Bosch, N. Morling, M. Muñoz-Nieto, E. Ortega-González, S. Pedrosa, R. Pérez, C. Solís, I. Yurrebaso, P. Gill, Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 47–54.
[6] C. Benschop, H. Haned, T. de Blaeij, A. Meulenbroek, T. Sijen, Assessment of mock cases involving complex low template DNA mixtures: a descriptive study, Forensic Sci. Int. Genet. 6 (2012) 697–707.
[7] H. Haned, C. Benschop, P. Gill, T. Sijen, Complex DNA mixture analysis in a forensic context: evaluating the probative value using a likelihood ratio model, Forensic Sci. Int. Genet. 16 (2015) 17–25.
[8] D.J. Balding, Evaluation of mixed-source, low-template DNA profiles in forensic science, PNASs 110 (30) (2013) 12241–12246.
[9] R.G. Cowell, T. Graversen, S.L. Lauritzen, J. Mortera, Analysis of forensic DNA mixtures with artefacts, Appl. Stat. 64 (1) (2015) 1–32.
[10] H. Haned, P. Gill, Analysis of complex DNA mixtures using the Forensim package, Forensic Sci. Int. Genet. Suppl. Ser. 3 (2011) e79–e80.
[11] E. Rykiel, Testing ecological models: the meaning of validation, Ecol. Model. 90 (1996) 229–244.
[12] D.C. Ince, L. Hatton, J. Graham-Cumming, The case for open computer programs, Nature 482 (2012) 485–488.
[13] D.K. Nordstrom, Models, validation, and applied geochemistry: issues in science, communication, and philosophy, Appl. Geochem. 27 (2012) 1899–1919.
[14] R.G. Sargent, Verification and validation of simulation models, J. Simulat. 7 (2013) 12–24.
[15] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013) 251–263.
[16] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, T. Caragine, Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int. Genet. 6 (6) (2012) 749–761.
[17] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (2011) 265–268.
[18] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, Forensic Sci. Int. Genet. 6 (2012) 762–774.
[19] J.M. Curran, P. Gill, M.R. Bill, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (2005) 47–53.
[20] K. Inman, N. Rudin, K. Cheng, C. Robinson, L. Kirschner, A. Inman-Semerau, K. Lohmueller, Lab Retriever: a software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles, BMC bioinformatics 16 (2015) 298.
[21] D. Balding, likeLTD: Likelihoods for Low-Template DNA Profiles, 2012.
[22] S. Grishechkin, K. Prokotfjena, Grape v.3.0, http://www.dna-soft.com/2014 (last retrieved on 20-1-).
[23] Scientific working group on DNA analysis methods, Guidelines for the validation of probabilistic genotyping systems, http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf2015 (retrieved 25/11/2015).
[24] B.A. McCarl, Model validation: an overview with some emphasis on risk models, Rev. Mark. Agr. Econ. 52 (1984) 153–173.
[25] L.N. Joppa, G. McInerny, R. Harper, L. Salido, K. Takeda, K. O'Hara, D. Gavaghan, S. Emmott, Troubling trends in scientific software use, Science 340 (2013) 814–815.

Forensic Science International: Genetics 22 (2016) 64–72

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

An assessment of the information content of likelihood ratios derived from complex mixtures

Clare D. Marsden[a], Norah Rudin[b], Keith Inman[c], Kirk E. Lohmueller[a,*]

[a] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, 621 Charles E. Young Dr. South, Los Angeles, CA 90095-1606, USA
[b] Forensic DNA Consultant, 650 Castro Street, Suite 120-404, Mountain View, CA, 94041, USA
[c] Department of Criminal Justice Administration, California State University, East Bay, 4069 Meiklejohn Hall, 25800 Carlos Bee Boulevard, Hayward, CA 94542, USA

ARTICLE INFO

Article history:
Received 22 September 2015
Received in revised form 6 January 2016
Accepted 16 January 2016
Available online 19 January 2016

Keywords:
Complex mixture
Likelihood ratio
Forensic DNA
Statistics
Simulation
Known non-contributor

ABSTRACT

With the increasing sensitivity of DNA typing methodologies, as well as increasing awareness by law enforcement of the perceived capabilities of DNA typing, complex mixtures consisting of DNA from two or more contributors are increasingly being encountered. However, insufficient research has been conducted to characterize the ability to distinguish a true contributor (TC) from a known non-contributor (KNC) in these complex samples, and under what specific conditions. In order to investigate this question, sets of six 15-locus Caucasian genotype profiles were simulated and used to create mixtures containing 2–5 contributors. Likelihood ratios were computed for various situations, including varying numbers of contributors and unknowns in the evidence profile, as well as comparisons of the evidence profile to TCs and KNCs. This work was intended to illustrate the best-case scenario, in which all alleles from the TC were detected in the simulated evidence samples. Therefore the possibility of drop-out was not modeled in this study. The computer program DNAMIX was then used to compute LRs comparing the evidence profile to TCs and KNCs. This resulted in 140,000 LRs for each of the two scenarios. These complex mixture simulations show that, even when all alleles are detected (i.e. no drop-out), TCs can generate LRs less than 1 across a 15-locus profile. However, this outcome was rare, 7 of 140,000 replicates (0.005%), and associated only with mixtures comprising 5 contributors in which the numerator hypothesis includes one or more unknown contributors. For KNCs, LRs were found to be greater than 1 in a small number of replicates (75 of 140,000 replicates, or 0.05%). These replicates were limited to 4 and 5 person mixtures with 1 or more unknowns in the numerator. Only 5 of these 75 replicates (0.004%) yielded an LR greater than 1,000. Thus, overall, these results imply that the weight of evidence that can be derived from complex mixtures containing up to 5 contributors, under a scenario in which no drop-out is required to explain any of the contributors, is remarkably high. This is a useful benchmark result on top of which to layer the effects of additional factors, such as drop-out, peak height, and other variables.

1. Introduction

As a consequence of both the increasing sensitivity of DNA typing methodologies, as well as mounting awareness by law enforcement of the perceived capabilities of DNA typing, complex mixtures consisting of DNA from two or more contributors are increasingly being encountered in forensic DNA profiles (N. Rudin and K. Inman, personal communication; [1–6]).

At least two factors may reduce the information content of multi-contributor samples as compared with single source samples. First, many of the possible alleles at a particular locus may be present in the evidence sample, diminishing the ability to exclude people as contributors to the mixture. Second, two or more contributors to the mixture may share the same alleles, increasing the difficulty of inferring the genotypes of the true contributors (TCs) of the mixture directly from the evidentiary sample. Together, these factors reduce the ability to distinguish TCs from known non-contributors (KNCs) in complex mixtures. These difficulties are exacerbated by forensic DNA evidence samples compromised by various conditions, such as low quantity and poor quality, that result in complex profiles exhibiting characteristics

* Corresponding author. Fax: +1 310 206 048.
E-mail address: klohmueller@ucla.edu (K.E. Lohmueller).

http://dx.doi.org/10.1016/j.fsigen.2016.01.008
1872-4973/

such as allelic drop-out, degradation, inhibition, peaks heights that do not reliably reflect the original contribution to the sample, and varying ratios of multiple contributors. In this work we focus on separating out the effects of multiple contributors.

Historically, binary approaches, such as combined probability of inclusion (CPI), and restricted or modified random match probability (RMP) have been used to estimate the evidential strength of mixed samples in forensic DNA analysis [7–10]. More recently, the likelihood ratio (LR) approach is gaining acceptance as a tool to estimate the weight of complex profiles [2,3,5]. The LR represents the ratio of probabilities of observing the alleles detected in an evidence profile under two mutually exclusive hypotheses, represented as the numerator (H_1) and denominator hypotheses (H_2). LR values greater than 1 are interpreted as indicating greater support for H_1 than H_2, whereas a LR less than 1 indicates greater support for H_2 than H_1 [11,12]. The standard mathematical depiction of the LR is:

$$\mathrm{LR} = \frac{\Pr(E|H_1)}{\Pr(E|H_2)}$$

Calculation of a LR requires specification of the total number of contributors, as well as the number of contributors meeting various conditions for both H_1 and H_2.

For situations encountered in forensic DNA, three categories of conditioned contributors are frequently encountered. The first category is an individual whose DNA is assumed present, usually because of the nature of the sample; this conditioned contributor is often categorized as "assumed." Contributors in this category are assumed to be present and therefore are conditioned contributors in both the numerator (H_1) and denominator (H_2). The second category is an individual for whom the weight of evidence is being assessed; this conditioned contributor is often characterized as a "suspected" or "hypothesized" contributor [13]. Contributors in this category follow different conditions in H_1 and H_2; typically this contributor is conditioned in H_1 and replaced with an unknown contributor in H_2. The third category is a contributor whose profile is unknown (unprofiled); unknown individuals are invoked to complete the total number of contributors.

Taking the simplest example, a single source sample, the numerator hypothesis would typically pose that the evidence derives from a single known individual (i.e. a profiled hypothesized contributor), whereas the denominator hypothesis replaces this known individual with an unknown (i.e. unprofiled) individual. In contrast, the hypotheses for mixtures expand to consider varying numbers of assumed, hypothesized and unknown contributors; thus, multiple pairs of competing hypotheses might be considered for a particular mixed sample. For example, under the assumption of a two person mixture, H_1 could posit that the evidence sample derives from one hypothesized and one assumed contributor, while the H_2 hypothesis might be that the evidence is explained by one unknown plus one assumed contributor. An alternative pair of hypotheses for the same mixture could be that under H_1 the mixture derives from one hypothesized contributor and one unknown contributor, while under H_2 the mixture derives from two unknown individuals.

Intuitively, we expect that a TC included in the numerator hypothesis should result in a $LR > 1$, indicating support for the proposition that the TC actually contributed to the sample. Conversely, we expect that a KNC assumed in the numerator hypothesis should result in a $LR < 1$, indicating support for the proposition that an unknown contributor is the TC to the sample. However, it has been shown that under certain scenarios these simplistic expectations fail. The earliest mention of this possibility surfaced when Evett [14] demonstrated that a two person mixture could yield a $LR < 1$ even when there existed confirmatory information for H_1, the numerator proposition (which Evett described as the 'prosecution proposition'). Much later, Brenner et al. [15] commented that altering the proposed number of contributors will change the LR from $LR > 1$ to $LR < 1$ when the hypothesized contributor carries the more common alleles in the mixture. Shortly thereafter, Weir et al. [8]. using the historical Polymarker® genetic typing kit on mixtures, showed that TCs may generate $LRs < 1$. Specifically, if all of the alleles at a particular locus were detected in the evidence profile, and the hypothesis in the numerator included at least one unknown contributor, the resulting LR could be less than 1 if the hypothesized contributor in the numerator carried common alleles at the locus. A small body of work suggests that, especially for mixtures, some non-trivial proportion of KNCs will generate $LRs > 1$ [16–19]. This is not only unsurprising, but statistically predicted. For example, when Gill et al. [16] proposed a method for measuring the robustness of an LR, they illustrated that simulated KNC profiles could produce $LRs > 1$. However, they only tested its usefulness on a handful of casework stains.

In spite of this earlier work, we are not aware of any published research that assesses how often these effects would be expected to occur in different types of mixtures, or to explore how different genotypes for the hypothesized contributor, might affect the results. In particular, the moderately variable loci typed in current short tandem repeat (STR)-based systems potentially give rise to the situations in which the LR for a TC included in the numerator hypothesis falls below 1, as well as those in which a KNC produces a $LR > 1$. Determining the frequency with which these effects occur, and under which particular circumstances, would add greatly to our understanding of the LRs produced for complex profiles.

As advances in technology began to allow laboratories to analyze challenging samples, it became immediately and abundantly clear that the community did not have the appropriate tools, nor the supporting research, to reliably interpret and weight the resulting complex profiles. Over the past few years, research on using LRs to assist in interpreting these profiles has produced publications on a number of issues, including the ability to estimate the number of contributors [20–25] and the effect of mis-specifying the number of contributors when computing LRs [24]. While Gill, et al. [16] proposed a method (implemented in LRMix [26]) to calculate the probability of a misleading LR, they incorporated the probability of both dropout and drop-in, but did not isolate the parameters of allele frequencies or number of contributors. We are not aware of any studies that systematically address how varying only the number of contributors and unknown contributors affects the ability to distinguish TCs from KNCs in mixed samples. Published studies [17,27,28] attempt either to separate the effects of multiple contributors from other variables, such as low template, drop-out, drop-in or peak heights, or focus on simpler hypotheses involving fewer contributors.

Unlike the LR, the CPI (aka random man not excluded, RMNE) does not require specification of the number of contributors in the sample. The lack of requirement to specify the number of contributors, combined with the ease of calculation, and perceived simplicity of explanation, has resulted in the widespread use and acceptance of this calculation. However, the CPI has been strongly criticized in the literature (reviewed in [1,7,29,30]) both because it discards information, and also because, in certain situations, it can be prejudicial to the hypothesized contributor for a variety of reasons [1,2,4,29–32].

Here, we aim to explore the capabilities and limitations of statistical approaches used to assess the strength of evidence derived from complex DNA mixtures based solely on the number of contributors (ranging from 2 to 5) and the frequency of alleles found in commonly-used STR loci. Nominally, we assessed how often and under what conditions TC generate $LRs < 1$ (termed

sensitivity by SWGDAM [33]) and KNCs generate LRs greater than 1 (termed specificity by SWGDAM [33]), and also how well TCs can be separated from KNCs in these mixtures by evaluating the distributions of LR values derived from each of these conditions. This latter metric is of critical practical importance, as it is the discrimination power (i.e. the degree of overlap) rather than the absolute LR values that inform us about the strength of evidence. Additionally, we explored how the variables of number of contributors and number of unknown contributors affect the ability to discriminate TCs from KNCs. Because complete profiles are not always recovered from forensic evidence samples, and because historical profiles generated from previous genetic analysis kits contain information at fewer loci, we also wanted to know how LRs computed from less informative profiles behave. To this end we performed our analyses on the loci found in the Identifiler® kit, as well a subset of 9 loci found in the Profiler Plus® kit. Finally, we wished to compare the behavior of two different kinds of statistical approaches commonly used in mixture analyses, the LR and the CPI, under the experimental scenarios detailed above.

Our results provide a best-case analysis of the information content of 2–5 person mixtures encountered in forensic DNA casework as we simulated data in which each of the contributors to a mixed sample was present in equal and sufficient amounts such that no drop-out is expected. We acknowledge that this is not a realistic casework scenario; most forensic evidence samples containing multiple contributors also tend to be of poor quality and low quantity, and the standard amount of DNA amplified using forensic DNA typing systems is often insufficient to completely represent multiple contributors. Nevertheless, it is not only useful, but requisite, to investigate a simplified model system so as to separate and understand the effects of different variables. Overall, our findings can be used as a benchmark for future analyses of more challenging samples involving other complex phenomena such as allelic drop-out or peak height differences.

2. Methods

2.1. Genotype simulation

Individual genotypes were simulated by sampling two alleles for each locus from a multinomial distribution with the parameters 2 and p, where p is the vector of allele frequencies for a specific locus. Essentially this process assumes Hardy–Weinberg equilibrium for each locus and linkage equilibrium across loci. For the 15-locus simulations, we considered the loci included in the Identifiler® kit. For 9-locus simulations we considered the subset of loci included in the Profiler Plus® kit. Allele frequency data from the Caucasian population, generated by the National Institutes of Standards and Technology (NIST) were used in all simulations [34].

2.1.1. Generation of mixtures

For each simulation replicate set, we first simulated six individual genotypes (C1, C2, C3, C4, C5, KNC). The first five individuals (denoted by "C") were next used as TCs to create the evidence mixtures and are hereafter referred to as C1, C2, C3, C4 and C5 (Table 1). Each set contained four mixtures [(C1,C2); (C1, C2,C3); (C1,C2,C3,C4); and (C1,C2,C3,C4,C5)] The sixth individual (KNC) in the set was simulated to represent a KNC (discussed below), and this genotype was never included in the mixtures. All mixtures were created assuming no drop-out (i.e. all of the alleles of the TC to the mixture were detected in the evidence profile).

2.1.2. Calculation of LRs

For each simulated mixture, we calculated LRs under two scenarios: one in which the hypothesized contributor (i.e. a contributor in the numerator) was the TC and one in which the hypothesized contributor was a KNC. In order to calculate the LR, the number of unknown and assumed contributors in the numerator (H_1) and denominator (H_2) hypotheses must be specified. However, for complex mixtures, many combinations of unknown and assumed contributors are possible. For example, for a two person mixture, H_1 could posit a mixture of DNA from one hypothesized contributor and one assumed contributor (e.g. the suspect and victim), and H_2 could specify one assumed contributor and one unknown contributor (e.g. the victim and an unknown individual). Alternatively H_1 could posit one hypothesized and one unknown contributor (e.g. suspect and one unknown contributor), and H_2 could specify two unknown contributors. In order to assess LRs derived from hypotheses with varying number of unknown contributors, we calculated LRs under 14 different hypotheses representing all possible combinations of conditioned contributors for each mixture (Table 1). For scenario 1, the C1 simulated for each replicate was always the hypothesized contributor in H_1 and thus never used in the denominator hypothesis. For scenario 2, in which a KNC is compared to the evidence profile instead of a TC, the simulated KNC was used as the hypothesized contributor in place of C1. We produced 10,000 replicate sets for each number of contributors (2–5) and set of hypotheses (see Table 1). This resulted in 140,000 LRs computed using a TC in the numerator, and

Table 1
Details of hypotheses investigated when calculating LRs for different mixtures.

Hypothesis[a]	Total # of contributors to the mixture	Contributors conditioned under H1[b]	Contributors conditioned under H2
h21	2	C1, C2	C2 + 1 UNK
h22	2	C1	2 UNK
h31	3	C1, C2, C3	C2, C3 + 1 UNK
h32	3	C1, C2	C2 + 2 UNK
h33	3	C1 + 2 UNK	3 UNK
h41	4	C1, C2, C3, C4	C2, C3, C4 + 1 UNK
h42	4	C1, C2, C3 + 1 UNK	C2, C3 + 2 UNK
h43	4	C1, C2 + 2 UNK	C2 + 3 UNK
h44	4	C1 + 3 UNK	4 UNK
h51	5	C1, C2, C3, C4, C5	C2, C3, C4, C5 + 1 UNK
h52	5	C1, C2, C3, C4 + 1 UNK	C2, C3, C4 + 2 UNK
h53	5	C1, C2, C3 + 2 UNK	C2, C3 + 3 UNK
h54	5	C1, C2 + 3 UNK	C2 + 4 UNK
h55	5	C1 + 4 UNK	5 UNK

[a] Hypotheses were named as follows: h[number of contributors in the mixture] [number of unknown contributors in H2]. In other words, h21 means [2 contributors to the mixture] [1unknown contributor in H2].

[b] C1 represents the hypothesized contributor, i.e. the conditioned contributor for whom the weight of evidence is being assessed. In order to calculate LR for a known non-contributor, C1 was replaced with KNC. UNK = an unknown contributor.

another 140,000 LRs using a KNC in the numerator, for a total of 280,000 LRs,

LR calculations for all hypotheses were conducted using DNAMIX v1.0 (http://genomine.org/dnamix/index.html). DNAmix input parameters were provided via a text file which was generated from the simulated genotype and mixture data with a custom python script. Similarly, the DNAmix output files were parsed with custom python scripts. For the KNC analysis, when the hypothesized (KNC) contributor carried alleles absent from the mixture at one or more loci, the LR was set to 0, because in such a situation, if drop-out is not possible, then the hypothesized contributor cannot be considered as a possible contributor.

2.1.3. Calculation of CPI

There exists both a need and natural curiosity to compare LRs with historical methods despite the differences in assumptions, variables considered, input data and fundamental approach of these two types of methods. To this end, a CPI was calculated for each mixture using the traditional formula, i.e., the square of the sum of the allele frequencies. The CPI specifies neither the number of individuals in the mixture nor the number of assumed and unknown individuals. Therefore, we calculated only one CPI for each mixture profile. Further, in order to perform a relevant comparison between the LRs computed above and the CPIs computed for the same mixture profiles, the CPI was inverted to 1/CPI to enable that comparison.

3. Results

Using simulations without allelic drop-out, we evaluated the strength of evidence that can be derived from complex mixtures of between 2 and 5 contributors, varying the numbers of unknown individuals in the proposed hypotheses (Table 1) and using either complete 15-locus or less informative 9-locus profiles.

3.1. Do true contributors (TCs) always yield LRs greater than 1?

A natural expectation for LRs which specify a TC in H_1 is that they will produce a result greater than one. For a complete 15-locus profile, we found that TCs resulted in LRs greater than one in 99.99% of replicates. However, in 7/140,000 simulations the TCs generated LRs less than 1 (Table 2). All 7 instances resulted from a 5-contributor mixture with at least one unknown contributor in H_1. Interestingly, the 7 instances were associated with just four replicates in the simulation (333, 855, 334[2×] and 6105[3×]), i.e., specific sets of simulated individual genotypes and the associated complex mixture (see Supplementary Table 1). Examination of the two replicates (6105 and 334) that produced LRs < 1 in multiple comparisons indicates that certain allelic combinations at a locus may predispose a profile to produce a LR less than 1 for a TC.

Specifically, for replicate 6105, the same seven loci (D16, D18, D21, D5, D7, FGA, vWA) produced LRs less than one for five person mixture hypotheses with 2, 3 and 4 unknown contributors (h52, h53, h54 see Table 1). For replicate 334, seven loci gave LRs < 1 (CSF, D19, D21, D5, FGA, D16 and D3), the first five of which were common to five person mixture hypotheses with 3 and 4 unknown contributors (h53, h54, see Table 1).

Still considering TCs in H_1, we further investigated how frequently the LR fell below 1 at individual loci. Our results show that for hypotheses including one or more unknown contributors

Table 3
Proportion of replicates generating LR < 1 for TC at 1, 2,3,4 and 5 or more loci by hypothesis based on a complicate 15 locus profile (see Table 1 for hypothesis notation).

	TC has LR < 1 at				
Hypothesis	1+ loci	2+ loci	3+ loci	4+ loci	5+ loci
h21	0	0	0	0	0
h22	44.15	10.12	1.49	0.17	0.04
h31	0	0	0	0	0
h32	77.22	40.95	15.64	4.01	0.73
h33	73.74	36.97	12.41	2.98	0.52
h41	0	0	0	0	0
h42	89.18	63.29	34.84	13.83	3.87
h43	89.65	62.55	33.04	13.15	4.05
h44	86.22	55.94	26.9	9.88	2.5
h51	0	0	0	0	0
h52	93.56	72.32	44.74	21.52	8.27
h53	95.65	79.28	52.93	28.31	11.39
h54	94.78	76.66	49.63	24.9	9.89
h55	93.44	72.73	43.83	21.09	7.4

Table 2
LR and 1/CPI values derived for different mixtures and hypotheses (see Table 1 for notation) for TC and KNC with a partial 15 locus profile.

	Percentage of replicates where TC LR			Percentage of replicates where KNC LR	
	>1	>1000	>1 million	>1	>1000
LR					
h21	100	100.00	100.00	0	0
h22	100	100.00	99.80	0	0
h31	100	100.00	99.98	0	0
h32	100	99.95	87.72	0	0
h33	100	99.80	63.96	0	0
h41	100	100.00	97.75	0	0
h42	100	98.23	54.36	0.01	0
h43	100	96.01	29.43	0.02	0
h44	100	93.47	15.97	0.02	0
h51	100	99.91	79.31	0	0
h52	99.99	90.69	28.03	0.09	0.01
h53	99.97	82.81	12.38	0.19	0.02
h54	99.97	75.38	6.00	0.21	0.01
h55	100	68.54	3.47	0.21	0.01
Total	99.99	93.20	55.58	0.05	0.004
1/CPI					
2 person	100	100.00	99.17	0	0
3 person	100	99.99	13.80	0	0
4 person	100	91.10	0.02	0	0.02
5 person	100	31.78	0.00	0.21	0
Total	100	23.06	8.07	0.015	0.001

in the numerator, 44% to 96% of simulations exhibited LRs < 1 at one or more loci (Table 3). Moreover, in mixtures with 5 contributors and one or more unknowns in the numerator (H_1; Table 3), 7–11% of simulations produced LRs less than one at 5 or more loci. Interestingly, although hypotheses with zero unknowns in H_1 never generated LRs less than 1, we found no evidence that increasing the number of unknowns increased the proportion of LRs less than one (Table 3). For example, in a 4-person mixture in which one, two or three unknowns were specified in H_1, the proportion of replicates in which one or more loci gave a LR less than one ranged from 86–89%. Conversely, the total number of contributors in the mixture greatly influences the proportion of TCs yielding LRs < 1 (Table 3).

3.2. Do known non-contributors (KNCs) always yield LRs less than 1?

A natural expectation for LRs which specify a KNC in H_1 is that they will produce a result less than one. Ideally, a discriminating typing system that performs well should yield LRs less than 1 for KNCs because such individuals are not actually present in the mixture. However, given the moderate heterozygosity of the STR loci, the varying number of alleles at each locus, and the distribution of allele frequencies, we expect some KNCs to yield LRs > 1 due to coincidental sharing of alleles. To quantify this effect, we determined the proportion of LRs in which a KNC posited in H_1 resulted in a LR greater than one. Indeed, in a small proportion of simulations (0.05%, 75/140,000, Table 2) the KNCs yielded LRs greater than one. In each of these instances, the result was associated with mixtures of 4 or 5 contributors, and in which one or more unknown was specified in H_1 (Table 2). In many instances, the LRs greater than one were only slightly larger than one. LRs greater than 1,000 were found in only 0.004% of comparisons (4/140,000 simulations; Table 2).

3.3. Comparison of the distributions of LRs between TCs and KNCs

To evaluate how well the LRs in this study can distinguish between TCs and KNCs, and to determine the effect of number of contributors and number of unknowns (Table 1), we plotted the distribution of LRs. Our comparisons are similar to the previously proposed Tippett plots [11,16,35,36]. For the mixtures and hypotheses considered in this study, TCs and KNCs produced largely distinct LR distributions (Fig. 1). This separation indicates that, even for 5 contributor samples for which mixture proportions are ignored and for which no drop-out is required to explain the hypothesized contributor(s), LRs produce reliable separation of TCs and KNCs. The distributions are most distinct for hypotheses in which 0 unknowns are posited for H_1. The distributions become wider and move closer to one as the number of contributors increases, and also as the number of unknowns increases. Interestingly, the dominant predictive variable appears to be the number of unknown contributors while the total number of contributors exhibits a lesser effect on the ability to separate TCs and KNCs.

3.4. Comparison to 9-locus profiles

The data presented so far are based on the loci included in a 15-locus profile. However, compromised samples often produce partial profiles, and of course historical profiles generated using,

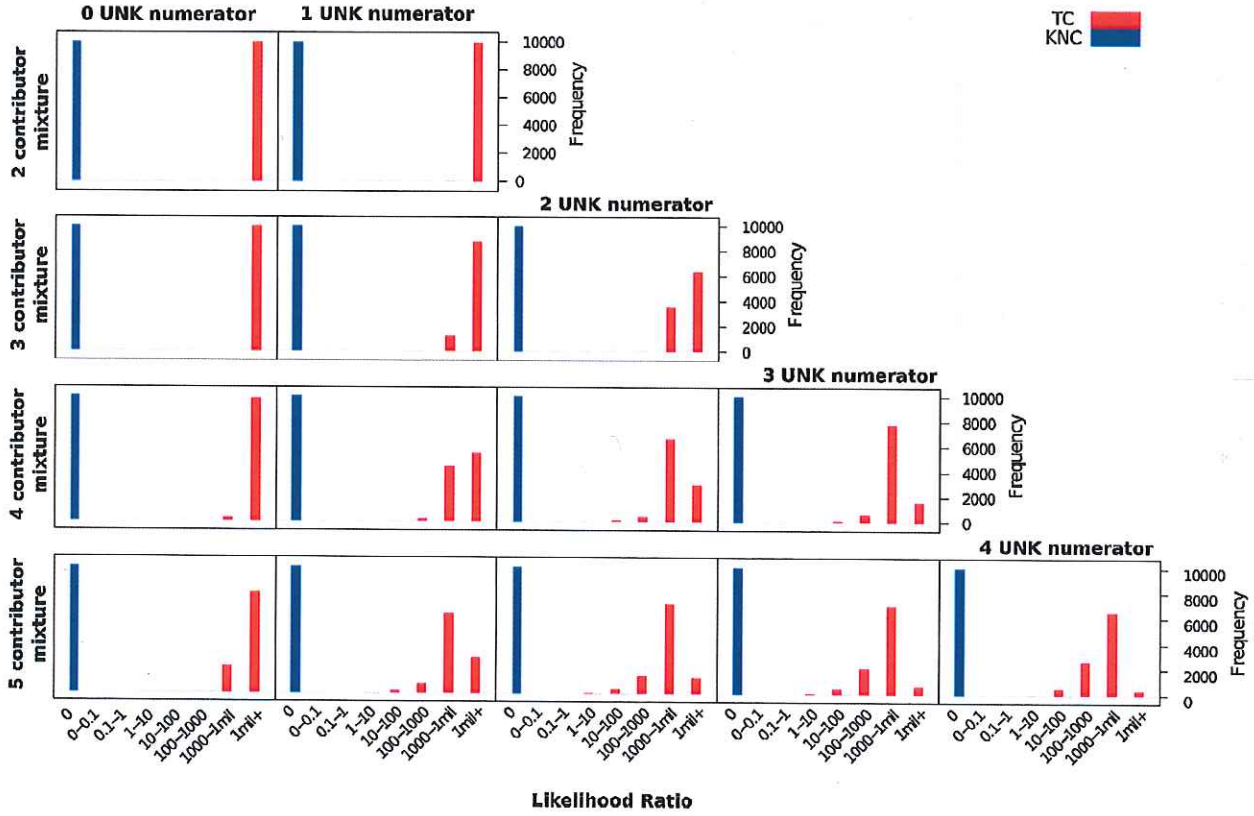


Fig. 1. Distribution of LRs for simulated mixtures. TC denotes true contributor (red) while KNC denotes a known non-contributor (blue). Rows denote the total number of contributors in the mixture while columns denote the number of unknowns in the numerator. The denominator always contains one addition unknown contributor. Overall, note the good separation of LRs between TCs and KNCs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for example, the Profiler Plus kit® comprise fewer loci. Therefore we also assessed LRs based on the 9 loci included in the Profiler Plus® kit. As expected, the 9-locus profiles generally produced smaller LRs for TCs, and larger LRs for KNCs (see Supplementary Table 2). Compared to the 15-locus profiles (Table 2), the proportion of simulation replicates for which TCs produced LRs greater than 1000 was notably smaller (71.72%) than for the 15-locus profiles (93.2%). Similarly, a smaller proportion of simulation replicates produced LRs greater than one million for TCs (25.40%) as compared to the 15-locus profiles (55.58%; Tables 2 and 3). Moreover, TCs producing LRs < 1 were an order of magnitude more common for the 9-locus profiles (64/140,000) than for the 15-locus profiles (7/140,000; Table 3). Finally, the 9-locus profiles in which H_1 specified a KNC produced LRs > 1 an order of magnitude more frequently (751/140,000; 0.54%) than did the 15-locus profiles (75/140,000 simulations; 0.05%). These results unsurprisingly indicate that profiles comprising fewer loci contain less information with which to distinguish TCs from KNCs. This pattern is reflected in the slightly greater overlap in the distributions of TCs and KNCs; however the separation was still relatively pronounced (Fig. 2).

3.5. LR vs. CPI

We next compared the LR statistics generated for each mixture profile with the CPI statistic. To enable a relevant comparison, we inverted the CPI to 1/CPI. It is mathematically impossible for 1/CPI values to fall below 1, therefore unlike the LR values, 1/CPI values were always greater than or equal to 1 (Table 2).

For TCs, LR values were generally larger than 1/CPI values (Table 2). This result is expected as, due to incomplete use of available data [7] and specifically when allelic drop-out is not required to explain the hypothesized contributor, the CPI typically underestimates the weight of the evidence. It is noteworthy that the difference between LR values and 1/CPI values was very small for 2 person mixtures, but considerably larger for 3, 4 and 5 person mixtures; this difference was exacerbated for hypotheses with fewer unknowns in H_1. For example, for 3 person mixtures, only 13% of 1/CPI values for TCs were greater than 1 million, in contrast to LR values, where 64–99% exceeded 1 million (Table 2).

For KNCs, 1/CPI values greater than 1 were found in a small number of replicates for 4 and 5 person mixtures. However, the magnitude of values that exceeded 1 was greater for 1/CPI than for LR values. In other words, when CPIs mis-identified a KNC as a contributor, the strength, hence inferred confidence, in the incorrect inference was greater.

4. Discussion

Complex mixtures are increasingly being encountered in forensic casework and therefore a need exists for robust statistical methods to accurately weigh the profiles generated from these samples. However, the presence of genotypes from multiple contributors may decrease the information content of complex mixtures. This results from the fact that, in samples derived from multiple contributors, a larger proportion of alleles in the population at a locus will be represented in a sample, thus reducing the power to exclude KNCs. Moreover, allele sharing

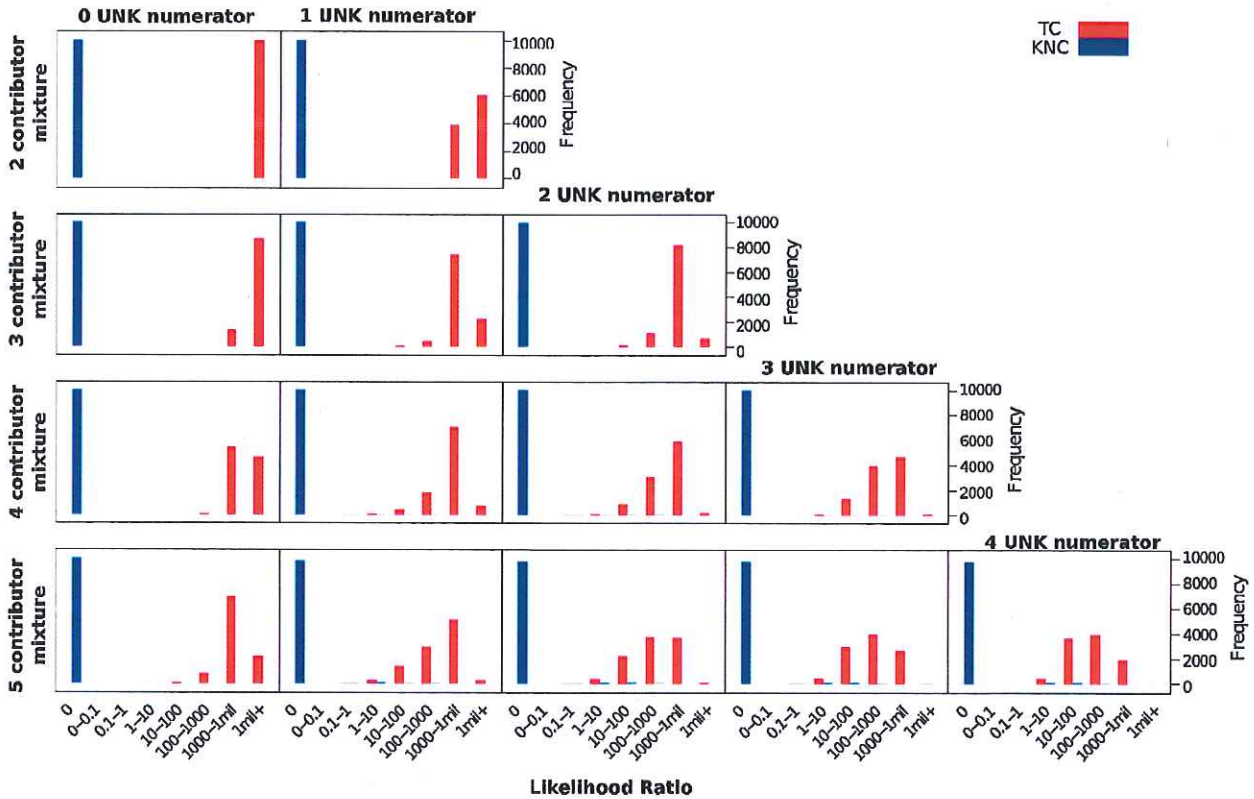


Fig. 2. Distribution of LRs for simulated mixtures including only 9 loci. TC denotes true contributor (red) while KNC denotes a known non-contributor (blue). Rows denote the total number of contributors in the mixture while columns denote the number of unknowns in the numerator. The denominator always contains one addition unknown contributor. Overall, the separation of LRs between TCs and KNCs is not as good as with 15 loci. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

between contributors makes it difficult to directly infer the genotypes of TC. The conditions under which it is possible to reliably distinguish TCs from KNCs (specificity) in samples with multiple contributors have not yet been fully characterized. In order to assess this in a controlled fashion, sets of Caucasian genotype profiles were simulated and used to create mixtures containing 2–5 contributors. All alleles from the contributor profiles were represented and no peak height information was used in computing LRs.

The distributions of LRs generated from TCs and KNCs for 15-locus profiles show good separation in complex mixtures, even in 5-person mixtures for which considerable allele sharing exists amongst the contributors (Fig. 1). Notably, the vast majority of TCs and KNCs are correctly inferred, even in multiple contributor profiles for which multiple unknowns are specified in the numerator hypothesis. Nevertheless, it is important to note that separation of TCs and KNCs is inversely proportional to both the number of contributors and the number of unknowns in the numerator hypotheses, indicating a decline in information content. While good separation of TCs and KNCs is achieved under the scenario in which full profiles of all contributors are represented, it is important to understand that this defines a best-case scenario. One type of information loss was simulated by reducing the number of loci to 9. Although TCs and KNCs still appeared relatively distinct, overlap increased due to the reduction in LR values for TCs and the increase in LR values for KNCs.

Therefore, for LRs for complex mixtures containing sufficient template to minimize the risk of drop-out, our present study fulfills the requirements of SWGDAM to study sensitivity and specificity [33]. However, further refinement of these recommendations is needed. In particular, the two guidelines that suggest determining "sensitivity" and "specificity" are as follows:

3.2.1.1. Sensitivity studies should demonstrate the potential for Type I errors (i.e., incorrect rejection of a true hypothesis), in which, for example, a contributor fails to yield a LR greater than 1 and thus his/her presence in the mixture is not supported."

3.2.2.1. Specificity studies should demonstrate the potential for Type II errors (i.e., failure to reject a false hypothesis), in which, for example, a non- contributor yields a LR greater than 1 and thus his/her presence in the mixture is supported.

Unfortunately, these definitions are at odds with the definitions of Type I and Type II errors in classical statistics [37]. Specifically, a Type I error is defined as a false positive, i.e. the sample is truly negative, but falsely identified as positive. This translates as a LR for a known non-contributor that incorrectly falls above one. Specificity, as typically defined in statistics, quantifies the ability of a test to *avoid* false positives. Type II error, on the other hand, is defined as a false negative, i.e. the sample is truly positive, but falsely identified as negative. This translates as a LR for a true contributor that incorrectly falls below 1. Again, as typically defined in statistics, sensitivity quantifies the ability to *avoid* false negatives. Thus the SWGDAM guidelines, as written would appear to have, at the very least, switched the nominal association of Type I and Type II errors; Type I errors should be associated with specificity, i.e. a known non-contributor that fails to yield an $LR < 1$; Type II errors should be associated with sensitivity, i.e. a true contributor that fails to yield an $LR > 1$. Additionally, specificity and sensitivity are, by definition, statistical measures of the performance of a binary classification test (e.g., disease present or absent). Using these types of statistical measures to quantify the performance of LRs, which, by definition, quantify the strength of the evidence for a particular proposition along a continuum of values, would seem overly simplistic at best. The reason for this is that any attempt to define a type I error is inextricably linked to the size of the observed LR. If the observed LR is close to one, many KNCs would be predicted to have a similar or larger LRs. This statistical uncertainty is not a problem because it is reflected in the LR itself. It may also be beneficial to examine the proportion of KNCs showing LRs greater than the observed value for the hypothesized contributor [26,38] rather than assigning an overall system-wide false-positive error rate. Nevertheless, in an attempt to categorize this work as per the SWGDAM guidelines as written, the simulations with TCs address guideline 3.2.1.1, quantifying the ability to measure false negatives, and the simulations with KNCs address guideline 3.2.2.1, quantifying the ability to measure false positives.

Previous studies have shown that complex mixtures can yield false negatives ($LRs < 1$ for TCs) and false positive ($LRs > 1$ for KNCs) results for complex mixtures. Previous work [8,14,15] showed that for systems comprising just a few specific loci, a TC may generate a $LR < 1$. However it was not clear whether the few low variability loci used were responsible for this phenomenon, or whether it would hold for an entire 15- or 9-locus profile of moderately variable loci. Here we show that TCs in fact do generate LRs less than 1 when using moderately polymorphic 15-locus STR profiles. However, this occurred only in a small number of cases (7/14,000 simulations) thus giving a misleading LR in just 0.05% of the mixtures. Indeed, the rarity of this result highlights the value of simulations, which permit very large number of replicates to be generated and thereby enabling the detection of rare events. These rare events were only associated with profiles of 5 contributors, and in which the numerator hypothesis included at least one unknown contributor (Table 2). Given that $LRs < 1$ for TCs occur more often with increasing numbers of unknown contributors in the mixture, combined with the discussion in Weir et al. [8]. we suggest that this effect may occur when the hypothesized contributor in the numerator accounts for the common alleles in the evidence profile. The remaining rarer alleles must then be explained by unknown contributors. However, in the denominator, an extra unknown contributor can be invoked to account for all of the alleles in the evidence. Because the extra unknown contributor can increase the probability of sampling the rarer alleles, the profile can produce a higher probability under the denominator hypothesis than under the numerator hypothesis, yielding an $LR < 1$.

While a composite 15-locus profile rarely generated a LR less than 1, we found that individual loci frequently generated LRs less than 1 for TCs (Table 3). This is an important practical consideration for forensic casework as it suggests that hypothesized contributors should not be excluded based solely on a LR less than 1 at a single locus. Rather, the LR for the entire profile should be used to inform decisions regarding whether an individual may be a contributor to a sample.

Previous work showed that KNCs generate LRs greater than 1 in a small percentage of samples [17–19]. Our work confirms this finding, showing that KNCs rarely generate misleading LRs, only at a rate of about 0.5%. In forensic casework, samples comprising multiple contributors frequently are additionally compromised by differential contributions and allelic drop-out. Thus, the information content will decrease further, and separation is expected to decline. For example, Mitchell et al. [17] examined low-template samples in which allelic drop-out was required to explain potential contributors. In contrast to our current results, this group observed LRs greater than 1 for KNCs even in 2 and 3 person mixtures (4 and 5 person mixtures were not assessed) and even assuming no unknowns in the numerator hypothesis.

The LR framework used in the current work does not model any factor that might complicate a DNA profile, such as drop-out, drop-in, degradation, inhibition, or contributor proportions. Thus, any KNC containing alleles absent from the mixture profile was automatically given a LR of 0. However, if drop-out were to be considered, these same individuals could potentially generate

LRs > 1. This highlights the ability to correctly support the exclusion of hypothesized contributors carrying discrepant alleles from high-level multi-contributor profiles in which drop-out is not expected. In low-level profiles, or for low-level contributors, the possibility of drop-out further compromises the ability to distinguish TCs from KNCs. In future work we will add allelic drop-out to further investigate the limits of separating TCs and KNCs in multiple contributor profiles in which the information content is further reduced due to missing information.

Another caveat to our study is that we did not incorporate peak height information in our calculations. If modeled correctly, peak heights could provide extra information by deconvolving the mixture [17,28,39,40] and in this regard would only improve the results presented here. For a profile with a discrepant ratio of contributors, when LRs calculated using an approach that models peak height are compared to LRs in which peak height information is completely ignored, higher LRs are, unsurprisingly, produced for TCs [27,28]. Peak height information can be incorporated in two ways: manually, prior to calculating an LR, and automatically modeled within the LR calculation. Various software tools offer a variety of different approaches to consider peak height information in an LR calculation [27,28,41]. While not a realistic casework scenario, our finding that, in a 5-person mixture profile in which complete information for all contributors is represented, the majority of simulations yield LRs greater than 1000 for TCs is telling. This result suggests that, for certain categories of profiles, a simple LR that does not model peak heights at all may be sufficient to provide convincing support for the presence of a TC. Additionally, if the person of interest is a minor contributor to a 5-person mixture, the stochastic effects of low template may tend to negate the benefits of incorporating allele peak heights. This information can assist laboratories in making important policy decisions. First, depending on the statistical tools available to a laboratory, they can define a complexity threshold above which they will decline to interpret a profile. Second, understanding which kinds of profiles truly benefit from a more complex treatment, and determining how often those profiles are encountered, is an important factor for laboratories to consider when weighing the cost of a more complex interpretational system as compared with simpler methods.

Historically, statistics such as the CPI have commonly been used to provide weight to mixture profiles. These types of statistics were easy to understand and calculate, and do not require the practitioner to specify the number of contributors. It has been argued that the CPI is easier to explain in court than LR approaches [9,10]. However, the literature overwhelmingly favors an LR approach, both because it has the ability to incorporate much more information, including peak heights, and also because it can model more complex variables such as drop-out and drop-in. Additionally, LRs address the question specifically relevant to the trier of fact by conditioning on the hypothesized contributor [7]. We compared the general performance of LRs and CPIs on the same profiles. We found that LRs and CPIs performed similarly on two person mixtures. However, for mixtures comprising 3 to 5 contributors, LRs generated larger values than CPIs for TCs (Table 2), particularly for hypotheses with fewer unknown contributors. This highlights the ability of LRs to use more information in the profile to provide stronger evidence in support of a TC. These data are consistent with previous studies [19,27,28] which suggest that the greater amount of information used by LRs provides a more accurate weight of evidence than CPIs, supporting the preferred use of LRs in mixture interpretation. Also consistent with previous work [7,24] is our finding that CPIs, which can never refute an inference of contribution by a specific donor, can more strongly support an incorrect inference of contribution than the comparable LR. In other words, CPIs have the potential to mislead more strongly, if not more often, than LRs.

5. Conclusion

Overall, these results imply that the weight of evidence that can be derived from complex mixtures containing up to 5 contributors, under a scenario in which no drop-out is required to explain any of the contributors, is remarkably high. This a useful benchmark result onto which future studies can layer the effects of additional variables, such as drop-out, contributor ratios, shared alleles, and other variables.

Author contributions

KEL, NR & KI conceived and designed the study, CDM conducted the simulations, CDM, KEL, NR and KI wrote the manuscript.

Acknowledgements

We thank an anonymous reviewer for comments that helped us clarify and improve parts of the manuscript. This work was supported by NIJ grant 2013-DN-BX-K029 to KPI and KEL.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2016.01.008.

References

[1] K.E. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (Suppl. 1) (2013) S243–249, doi: http://dx.doi.org/10.1111/1556-4029.12017.
[2] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, et al., DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101, doi:http://dx.doi.org/10.1016/j.forsciint.2006.04.009.
[3] P. Gill, L. Gusmao, H. Haned, W.R. Mayr, N. Morling, W. Parson, et al., DNA commission of the International Society of Forensic Genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Sci. Int. 6 (2012) 679–688, doi:http://dx.doi.org/10.1016/j.fsigen.2012.06.002.
[4] H. Kelly, J.-A. Bright, J.S. Buckleton, J.M. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles, Sci. Justice 54 (2014) 66–70, doi:http://dx.doi.org/10.1016/j.scijus.2013.07.003.
[5] C.D. Steele, D.J. Balding, Statistical evaluation of forensic DNA profile evidence, Annu. Rev. Stat. Its Appl. 1 (2014) 361–384, doi:http://dx.doi.org/10.1146/annurev-statistics-022513-115602.
[6] N. Gilbert, Science in court: DNA's identity crisis, Nature 464 (2010) 347–348, doi:http://dx.doi.org/10.1038/464347a.
[7] J. Buckleton, J. Curran, A discussion of the merits of random man not excluded and likelihood ratios, Forensic Sci. Int. 2 (2008) 343–348, doi:http://dx.doi.org/10.1016/j.fsigen.2008.05.005.
[8] B.S. Weir, C.M. Triggs, L. Starling, L.I. Stowell, K.A. Walsh, J. Buckleton, Interpreting DNA mixtures, J. Forensic Sci. 42 (1997) 213–222.
[9] B. Budowle, A.J. Onorato, T.F. Callaghan, A. Della Manna, A.M. Gross, R.A. Guerrieri, et al., Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, J. Forensic Sci. 54 (2009) 810–821, doi:http://dx.doi.org/10.1111/j.1556-4029.2009.01046.x.
[10] D.N.A.A.M. Scientific Working Group on, SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories, (2010). http://www.fbi.gov/about-us/lab/codis/swgdam.pdf (accessed 1.1.2010).
[11] I.W. Evett, B.S. Weir, Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists, Sinauer Associates, Sunderland, MA, 1998.
[12] J. Buckleton, A Framework for interpreting evidence, in: J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, Boa Raton, FL, 2005, pp. 27–63.
[13] K. Inman, N. Rudin, K. Cheng, C. Robinson, A. Kirschner, L. Inman-Semerau, et al., Lab Retriever: a software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles, BMC Bioinformatics 16 (2015) 298, doi:http://dx.doi.org/10.1186/s12859-015-0740-8.

[14] I.W. Evett, On meaningful questions: a two-trace transfer problem, J. Forensic Sci. Soc. 27 (1987) 375–381, doi:http://dx.doi.org/10.1016/S0015-7368(87)72785-6.
[15] C.H. Brenner, R. Fimmers, M.P. Baur, Likelihood ratios for mixed stains when the number of donors cannot be agreed, Int. J. Legal Med. 109 (1996) 218–219, doi:http://dx.doi.org/10.1007/BF01225523.
[16] P. Gill, J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, et al., Interpretation of complex DNA profiles using empirical models and a method to measure their robustness, Forensic Sci. Int. 2 (2008) 91–103, doi:http://dx.doi.org/10.1016/j.fsigen.2007.10.160.
[17] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, et al., Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int. 6 (2012) 749–761, doi:http://dx.doi.org/10.1016/j.fsigen.2012.08.007.
[18] K.E. Lohmueller, N. Rudin, K. Inman, Analysis of allelic drop-out using the identifiler and PowerPlex 16 forensic STR typing systems, Forensic Sci. Int. Genet. 12C (2014) 1–11, doi:http://dx.doi.org/10.1016/j.fsigen.2014.04.003.
[19] M.W. Perlin, K. Dormer, J. Hornyak, L. Schiermeier-Wood, S. Greenspoon, TrueAllele casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases, PLoS One 9 (2014) e92837, doi:http://dx.doi.org/10.1371/journal.pone.0092837.
[20] T. Egeland, I. Dalen, P.F. Mostad, Estimating the number of contributors to a DNA profile, Int. J. Legal Med. 117 (2003) 271–275, doi:http://dx.doi.org/10.1007/s00414-003-0382-7.
[21] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, J. Forensic Sci. 50 (2005) 1361–1366.
[22] H. Haned, L. Pène, F. Sauvage, D. Pontier, The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci. Int. Genet. 5 (2011) 281–284, doi:http://dx.doi.org/10.1016/j.fsigen.2010.04.005.
[23] H. Haned, L. Pène, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci. 56 (2011) 23–28, doi:http://dx.doi.org/10.1111/j.1556-4029.2010.01550.x.
[24] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci. Int. Genet. 1 (2007) 20–28, doi:http://dx.doi.org/10.1016/j.fsigen.2006.09.002.
[25] H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping, Forensic Sci. Int. Genet. 16 (2015) 172–180, doi:http://dx.doi.org/10.1016/j.fsigen.2014.11.010.
[26] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013) 251–263, doi:http://dx.doi.org/10.1016/j.fsigen.2012.11.002.
[27] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PloS One 4 (2009) e8327, doi:http://dx.doi.org/10.1371/journal.pone.0008327.
[28] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele(R) DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447, doi:http://dx.doi.org/10.1111/j.1556-4029.2011.01859.x.
[29] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. 4 (2009) 1–10, doi:http://dx.doi.org/10.1016/j.fsigen.2009.03.003.
[30] J.M. Curran, J. Buckleton, Inclusion probabilities and dropout, J. Forensic Sci. 55 (2010) 1171–1173, doi:http://dx.doi.org/10.1111/j.1556-4029.2010.01446.x.
[31] P. Gill, R.M. Brown, M. Fairley, L. Lee, M. Smyth, N. Simpson, et al., National recommendations of the Technical UK DNA working group on mixture interpretation for the NDNAD and for court going purposes, Forensic Sci. Int. 2 (2008) 76–82, doi:http://dx.doi.org/10.1016/j.fsigen.2007.08.008.
[32] P. Gill, J. Buckleton, A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number, Forensic Sci. Int. 4 (2010) 221–227, doi:http://dx.doi.org/10.1016/j.fsigen.2009.09.008.
[33] Scientific Working Group on DNA Analysis Methods, Guidelines for the validation of probabilistic genotyping systems, (2015). http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf.
[34] J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, M.C. Kline, Allele frequencies for 15 autosomal STR loci on U.S. Caucasian African American, and Hispanic populations, J. Forensic Sci. 48 (2003) 908–911.
[35] I.W. Evett, J. Scranage, R. Pinchin, An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science, Am. J. Hum. Genet. 52 (1993) 498–505.
[36] I.W. Evett, P.D. Gill, J.K. Scrange, B.S. Weir, Establishing the robustness of short-tandem-repeat statistics for forensic applications, Am. J. Hum. Genet. 58 (1996) 398–407.
[37] R. Sokal, F.J. Rohlf, Biometry, 4th edition, W. H. Freeman, New York, 2011.
[38] G. Dørum, Ø. Bleka, P. Gill, H. Haned, L. Snipen, S. Sæbø, et al., Exact computation of the distribution of likelihood ratios with forensic applications, Forensic Sci. Int. Genet. 9 (2014) 93–101, doi:http://dx.doi.org/10.1016/j.fsigen.2013.11.008.
[39] T.M. Clayton, J.P. Whitaker, R. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Sci. Int. 91 (1998) 55–70, doi:http://dx.doi.org/10.1016/S0379-0738(97)00175-8.
[40] T. Wang, N. Xue, J.D. Birdwell, Least-square deconvolution: a framework for interpreting short tandem repeat mixtures, J. Forensic Sci. 51 (2006) 1284–1297, doi:http://dx.doi.org/10.1111/j.1556-4029.2006.00268.x.
[41] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528, doi:http://dx.doi.org/10.1016/j.fsigen.2013.05.011.

From: [redacted]
To:
Subject: References on Hair
Date: Wednesday, December 14, 2016 9:09:47 PM

Dear Eric,

Greetings from China. This is Henry Lee. I has just receive a copy of AAFS newsletter with information about you are looking for relevant articles about forensic examination of hair. Enclosed is a list of my publications regards hair examination for your file.

Best wishes,

Dr. Henry Lee

1. Forensic Hair Examination. Lee, H.C. and DeForest, P.R., in C.H. Wecht (ed.) Forensic Sciences, Vol. 2, Matthew Bender, New York, 1984
2. Forensic Examination of Hair Evidence, Henry Lee & Elaine Pagliaro, in Forensic Sciences, Lexis-Nexis, 2010.
3. Lee, H.C. et al, "The Reliability of ABO Grouping of Human Hair", AAFS Abstract, Bll, 1983.
4. Lee, H.C., Mills, R.J., Settachatgul, K., "Forensic Examination of Human Hair", 4th Indo-Pacific Congress on Legal Medicine and Forensic Sciences, Bangkok, Thailand, November 2, 1992, (abstract)
5. Physical Evidence in Forensic Science 2nd Edition, Henry Lee & Howard Harris, Lawyers & Judges Publishing, 2006

From: [redacted]
To:
Subject: CTS Report Submission for PCAST
Date: Tuesday, December 13, 2016 5:23:40 PM
Attachments: CTS_SummaryReports_Firearms.zip
CTS_SummaryReports_Imprints.zip

PCAST Co-Chair Eric Lander,

Collaborative Testing Services, Inc, (CTS) is an independent proficiency test provider with 40 years of experience supplying discipline appropriate samples to forensic laboratories worldwide. The recent request by the President's Council of Advisors on Science and Technology (PCAST) to provide information to assist in the establishment of foundational validity and estimation of accuracy was reviewed by our company and technical staff, and we believe our reports offer substantial information that can advance your endeavors.

From the supplied list of forensic feature comparison methods, CTS has extensive historical information regarding two of these disciplines. We have been producing Firearms Analysis tests since 1978 and Footwear Analysis tests since 1985. We respectfully submit CTS Summary Reports for these disciplines as follows:

Firearms Analysis:

- 31 reports from 2001 through 2016, which represents two tests per year, with only one from 2016 as our final test of the year has not yet been published.
- These reports contain information from a yearly average of 450 examiners from 175 laboratories in 30 countries.

Footwear Analysis (CTS Imprint/Impression Test):

- 16 reports from 2001 through 2016, which represents one test per year.
- These reports contain information from an average of 250 examiners from 150 laboratories in 20 countries.

Within the PCAST report "Forensic Science in the Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods.", the authors stressed the use of Black Box studies as a key tool in establishing foundational validity and estimating the accuracy of these methods. CTS tests are very similar to Black Box studies in that the examiners do not know the expected answer, they all receive the same input (samples) and their output (results) are compared to the results of the other participants. While CTS focuses on consensus results when highlighting inconsistencies, the ground truth of the supplied samples is known and presented at the beginning of the report within our Manufacturer's Information statement.

CTS believes that these reports support the foundational validity of these practices. Within both disciplines, an average consensus of > 95% of participants match the ground truth in 46 of 47 reports. It is important to note that CTS tests do not require examiners to evaluate the samples for value prior to making comparison decisions as in other studies, and examiners are limited to the choices: "yes", "no", "inconclusive" for their comparison decisions. These are just a few of the factors that should be considered before utilizing our data sets as an accuracy estimation for a method.

We would be happy to answer any questions concerning CTS Reports.

Regards,

Catherine Brown
Vice President, Operations
Collaborative Testing Services, Inc.
[redacted]
www.collaborativetesting.com
www.ctsforensics.com

Documents received from Collaborative Testing Services

1. Test No 01-526: Firearms Examination
2. Test No 01-527: Firearms Examination
3. Test No 02-526: Firearms Examination
4. Test No 02-527: Firearms Examination
5. Firearms Examination Test No 03-526 Summary Report
6. Firearms Examination Test No 03-527 Summary Report
7. Firearms Examination Test No 04-526 Summary Report
8. Firearms Examination Test No 04-527 Summary Report
9. Firearms Examination Test No 05-526 Summary Report
10. Firearms Examination Test No 05-527 Summary Report
11. Firearms Examination Test No 06-526 Summary Report
12. Firearms Examination Test No 06-527 Summary Report
13. Firearms Examination Test No 07-526 Summary Report
14. Firearms Examination Test No 07-527 Summary Report
15. Firearms Examination Test No 08-526 Summary Report
16. Firearms Examination Test No 08-527 Summary Report
17. Firearms Examination Test No 09-526 Summary Report
18. Firearms Examination Test No 09-527 Summary Report
19. Firearms Examination Test No 10-526 Summary Report
20. Firearms Examination Test No 10-527 Summary Report
21. Firearms Examination Test No 11-526 Summary Report
22. Firearms Examination Test No 11-576 Summary Report
23. Firearms Examination Test No 12-526 Summary Report
24. Firearms Examination Test No 12-527 Summary Report
25. Firearms Examination Test No 13-526 Summary Report
26. Firearms Examination Test No 13-527 Summary Report
27. Firearms Examination Test No 14-526 Summary Report
28. Firearms Examination Test No 14-527 Summary Report
29. Firearms Examination Test No 15-526 Summary Report
30. Firearms Examination Test No 15-527 Summary Report
31. Firearms Examination Test No 16-526 Summary Report
32. Test No 01-533 Imprint/impression Evidence
33. Test No 02-533 Imprint/impression Evidence
34. Imprint/impression Evidence Test No 03-533
35. Imprint/impression Evidence Test No 04-533 Summary Report
36. Imprint/impression Evidence Test No 05-533 Summary Report
37. Imprint/impression Evidence Test No 06-533 Summary Report
38. Imprint/impression Evidence Test No 07-533 Summary Report
39. Imprint/impression Evidence Test No 08-533 Summary Report
40. Imprint/impression Evidence Test No. 09-533 Summary Report
41. Imprint/impression Evidence Test No 10-533 Summary Report

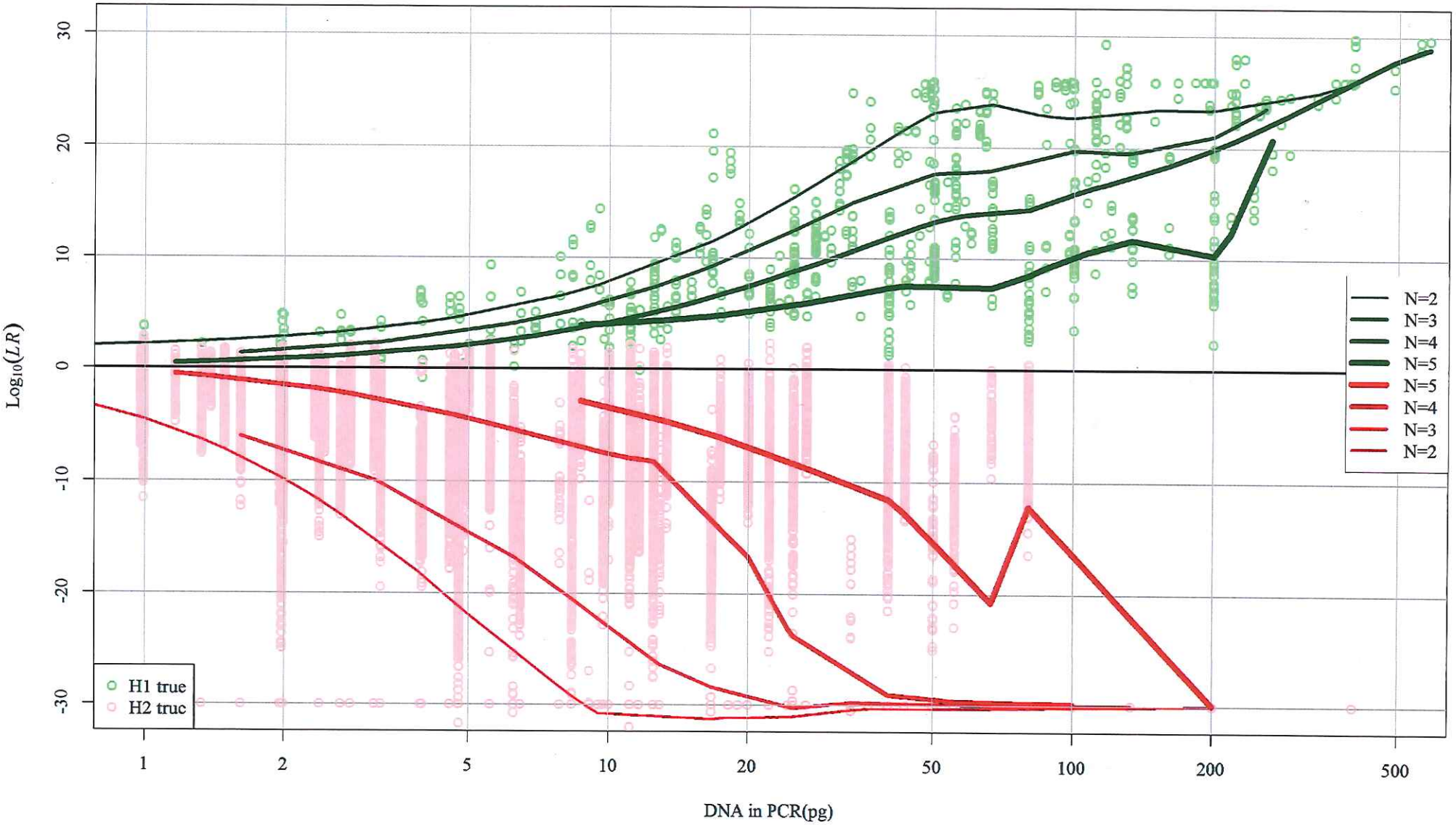
42. Imprint/impression Evidence Test No 11-533 Summary Report
43. Imprint/impression Evidence Test No 12-533 Summary Report
44. Imprint/impression Evidence Test No 13-533 Summary Report
45. Imprint/impression Evidence Test No 14-533 Summary Report
46. Imprint/impression Evidence Test No 15-533 Summary Report
47. Imprint/impression Evidence Test No 16-533 Summary Report

From: [redacted]
To:
Subject: Fwd: Results for 5p testing
Date: Friday, January 6, 2017 9:32:58 AM

From John Buckleton.

---------- Forwarded message ----------
From: **John Buckleton** <[redacted]
Date: Fri, Dec 9, 2016 at 4:43 AM
Subject: Results for 5p testing
To: [redacted]

Eric, attached are the results from Duncan Taylor for the 5p testing. Please feel free to distribute within PCAST. John


30
20
10
0
-10
-20
-30
Log10(LR)
1
2
5
10
20
50
100
200
500
DNA in PCR(pg)
N=2
N=3
N=4
N=5
N=5
N=4
N=3
N=2
H1 true
H2 truc

From: Grant, Pat
To: FN-OSTP-PCAST
Subject: hair paper
Date: Wednesday, December 14, 2016 4:46:22 PM

This email is in response to the request in an AAFS news alert. I presume the following paper was not considered by the PCAST report (altho don't know for sure). It's likely completely ignorable for that objective, but you can make the decision. It was refereed and published in the classified literature, and was well outside the mainstream of forensic hair analysis. The technique does work, however, and the classification was only OUO:

"Forensic Analysis of Hair for Inorganic and Actinide Signature Species Indicative of Nuclear Proliferation"
Journal of Intelligence Community Research & Development, Intelink, 10 pp, 2009.
P.M. Grant, A.M. Volpe, and K.J. Moody

Pat Grant
Forensic Science Center
LLNL, L-091
Livermore, CA 94550
[redacted]

Forensic Science International 266 (2016) 399–411

Contents lists available at ScienceDirect

Forensic Science International

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

Technical Note

Quantifying randomly acquired characteristics on outsoles in terms of shape and position

Jacqueline A. Speir [a,*], Nicole Richetelli [a], Michael Fagert [a,1], Michael Hite [a], William J. Bodziak [b]

[a] West Virginia University, 208 Oglebay Hall, PO Box 6121, Morgantown, WV 26506, United States
[b] Bodziak Forensics, 38 Sabal Bend, Palm Coast, FL 32137, United States

ARTICLE INFO

Article history:
Received 19 December 2015
Received in revised form 7 June 2016
Accepted 8 June 2016
Available online 23 June 2016

Keywords:
Footwear
Shoeprints
Randomly acquired characteristics
Accidentals
Fourier descriptors
Feature vectors

ABSTRACT

Footwear evidence has tremendous forensic value; it can focus a criminal investigation, link suspects to scenes, help reconstruct a series of events, or otherwise provide information vital to the successful resolution of a case. When considering the specific utility of a linkage, the strength of the connection between source footwear and an impression left at the scene of a crime varies with the known rarity of the shoeprint itself, which is a function of the class characteristics, as well as the complexity, clarity, and quality of randomly acquired characteristics (RACs) available for analysis. To help elucidate the discrimination potential of footwear as a source of forensic evidence, the aim of this research is to further characterize the chance association in position, shape, and geometry of RACs on a semi-random selection of footwear. To accomplish this goal in an efficient manner, a partially automated image processing chain was required, including steps for automated feature characterization. This technical note details the methods, procedures, and type of results available for subsequent statistical analysis after processing a collection of more than 1000 shoes and 57,426 randomly acquired characteristics.

1. Introduction

Although footwear impression evidence can provide a wealth of information about a crime, including potential suspects, the total number of possible offenders, and the most probable series of events associated with a reconstruction, this evidence is often undervalued (or even overlooked) due to limited knowledge about how to collect, analyze, and interpret footwear impressions [1]. Part of the reason for this disconnect may be the difficulty associated with collecting sufficient-sized and community-shared databases for extensive research and study, which would allow the legal and forensic community to fully appreciate the value of this type of evidence. The fact is, footwear research is extremely time-consuming and labor intensive, regardless of whether the analyst is interested in characterizing class, randomly acquired characteristics (RACs), or both. Although class features hold incredible value, this project deliberately disregards class characteristics and instead focuses on RACs or accidental features such as nicks, tears, holes, and cuts that typically develop on outsoles as a function of wear. The reason for this narrow focus in scope is primarily four-fold. First, class features have received some research attention in the past [2–11] and this trend is likely to continue in the future. As a result, this investigative effort intentionally sought out the less-traveled parallel track concerning characterization of accidental features, while simultaneously collecting sufficient data to allow for subsequent class analysis downstream. This motivation was largely driven by the fact that the majority of existing RAC databases are limited in terms of statistical size, and typically restricted to less than 50–100 shoes [12–17], save two exceptions known to the authors with sample sizes larger than 500–800 paired impressions [18–20]. As such, this research will help alleviate the scarcity of statistical information concerning accidental features by contributing information on more than 1000 impressions and 57,426 randomly acquired characteristics. Second, the National Academy of Sciences' (NAS) 2009 report on *Strengthening Forensic Science in the United States* encouraged studies to shed light on the variability of randomly acquired characteristics, including relative frequency of features, and the appropriate use of statistical standards [21]. Third, the Scientific Working Group for Shoeprint and Tire Tread Evidence

* Corresponding author. Tel.: +1 304 293 9233.
E-mail address: Jacqueline.Speir@mail.wvu.edu (J.A. Speir).
[1] Current address: Kansas City Missouri Police Crime Laboratory, 2645 Brooklyn Avenue, Kansas City, MO 64127, United States.

http://dx.doi.org/10.1016/j.forsciint.2016.06.012
0379-0738/

(SWGTREAD) requested focused research in the area of "*Random Placement Shape and/or Placement of Randomly Acquired Characteristics*" [22], and finally, SWGTREAD also requested intensified research in the area of "*Mathematical Probabilities of Randomly Acquired Characteristics*" [23]. Given these challenges, the first goal (and bottleneck) of this project was data acquisition. The remainder of this technical note describes the manner in which more than 1000 worn shoes (obtained from a variety of sources including personal donations, corporate donations, and purchases from local thrift stores) were sequentially processed via a combination of automated and user-fed algorithms allowing for identified RACs to be extracted and characterized in terms of shape, geometry, and physical location.

2. Material and methods

Available defining characteristics associated with more than 1000 shoes have been recorded, including make, model, size, manufacturer product code, degree of wear, and the presence of either microcellular material or Schallamach patterns as detailed in Tables 1–6. As necessary, each shoe was gently washed (using warm water) to remove debris (*i.e.*, this research does not account for the possible presence of transient RACs, such as rocks, gum, etc.). When dry, each outsole was scanned at 600PPI with an Epson Expression 11000XL Graphic Arts Scanner. Post-outsole scanning, Handiprint exemplars were created [1] using a Zephyr® brush (A-1-0200 Arrowhead Forensics, trimmed to a total length of approximately 1 inch), Lightning® Black Powder (1-4005 CSI Forensic Supply) and Handiprint sheets with clear polyester covers (2-3150 CSI Forensic Supply). To create each exemplar, the Handiprint sheet was prepared by removing the clear polyester sheet and allowing the flexible Handiprint material to rest (reform shape, adhesive side-up) while lightly dusting the outsole with the powder and Zephyr® brush.

Table 1
Frequency of shoe type.

Type	Number
Athletic	838
Dress shoe	88
Boot	56
Sandal	18
Total	1000

Table 2
Degree of wear. Shoes with light wear have discernible texture. Shoes with moderate wear may show some bald spots and lost texture. Shoes with heavy wear have a near complete loss of texture, many or large bald spots, and possible holes or areas where the outsole has worn away.

Wear	Number
Light	281
Moderate	456
Heavy	263
Total	1000

Table 3
Presence of microcellular material on the outsole.

Microcellular material	Number
Present	108
Absent	892
Total	1000

Table 4
Presence of Schallamach pattern on the outsole.

Schallamach pattern	Number
Present	743
Absent	257
Total	1000

Table 5
Frequency of manufacturer/brand.

Manufacturer/brand	Number
Adidas	28
Asics	30
Brooks	10
Converse	30
Hoka	36
New balance	20
Nike	294
Puma	14
Reebok	160
Skechers	12
Under armour	60
Unknown	26
Other (fewer than 10 shoes)	280
Total	1000

Table 6
Frequency of men's and women's shoe sizes. Note: shoes of unknown size account for the remaining 106 shoes (approximately 10%) of the database. Please note that size includes the full and half size; for example, a size 6 includes size 6 and size 6.5.

Men's size	Number	Women's size	Number
Size 5	2	Size 4	4
Size 6	4	Size 5	2
Size 7	28	Size 6	10
Size 8	54	Size 7	56
Size 9	148	Size 8	70
Size 10	200	Size 9	46
Size 11	162	Size 10	22
Size 12	62	Size 11	8
Size 13	14	Size 12	2
Total	674	Total	220

During powder application, the outsole was brushed in at least three directions; North-South (toe/heel), East-West (medial/lateral) and diagonally to ensure full coverage. After dust application, the shoe was tapped three-four times to dislodge excess dust, before placing the outsole on top of the prepared Handiprint sheet sitting on the laboratory benchtop. The Handiprint + shoe combination was slowly pulled off of the benchtop toward the analyst, while the researcher used his or her hands to gently add pressure on the non-adhesive side of the Handiprint (pressing the outsole against the tacky side of the Handiprint to maximize tight contact). When the Handiprint + - shoe was fully removed from the laboratory benchtop, the analyst then used a paper towel or fingerprint roller to gently reapply pressure between the Handiprint and outsole to again maximize contact. When complete, the Handiprint was pulled from the outsole and laid flat on the benchtop. The clear polyester cover was then slowly re-applied from bottom to top in a type of rastering process to minimize the introduction of air pockets between the Handiprint and protective cover. After development, the Handiprint was likewise scanned at 600PPI. Both are illustrated in Fig. 1 for a size 9 men's Converse Chuck Taylor® All Star® with moderate wear and Schallamach patterns.

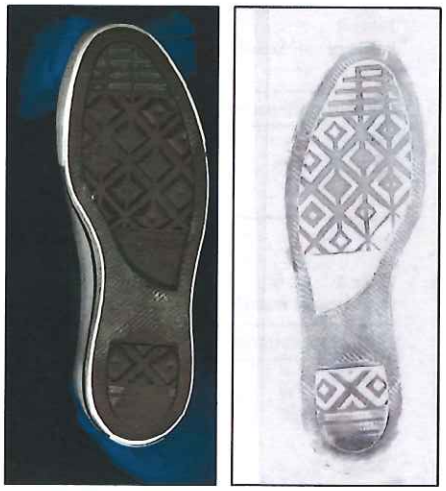


Fig. 1. Example of outsole (left) and Handiprint exemplar (right) scans.

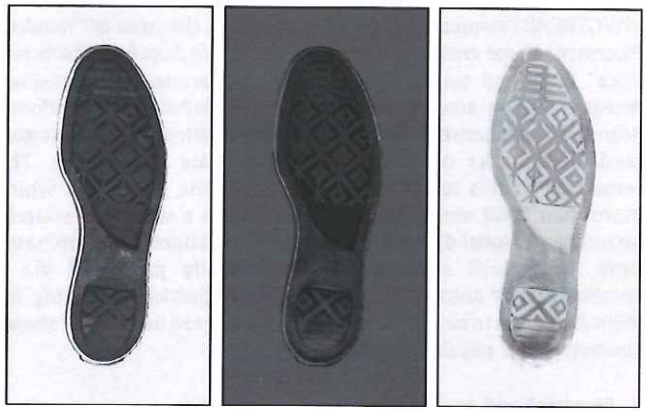


Fig. 2. Registered and background subtracted outsole scan (left) and Handiprint scan (right). The middle image is an overlay of the outsole and Handiprint illustrating co-registration.

In order to facilitate the automated downstream extraction of RAC shape and position, the outsole and exemplar were background subtracted and registered using identified control points. This process required the analyst to identify eight common geometric shapes that were patent on both the outsole and the exemplar. The features selected for registration varied per shoe, but needed to be distributed as evenly as possible around the perimeter of the outsole (a minimum of two on the toe, two on the heel, and the remaining four on the lateral and medial sides of the shoe) and generally consisted of class characteristics with sharp boundaries, such as corners in polygonal-geometric shapes (and lettering in logos, if applicable).

To expedite this process, a simple graphical user interface was constructed that opened two paired images. The first consisted of the scanned version of the outsole, while the second displayed the mirrored version of the Handiprint exemplar scan. With both images in a common orientation, the analyst used the cross-hair of the cursor on his or her mouse to designate mated-points between the images (open windows). Using this process, any number of mated points could have been selected, but as a compromise in terms of efficiency and accuracy, eight total ground control points were selected. Of the two possible images to use as a base, the outsole was selected, which meant during transformation, the Handiprint exemplar was translated, rotated, and scaled (as necessary) to bring it into registration with the outsole. This transformation was performed using a first order polynomial with least-squares fitting (note that a first order polynomial was selected over an affine transformation in order to handle slight shearing in the toe and heel that is not uncommon when creating Handiprint exemplars).

In addition to this co-registration, the background (non-tread areas) of both the outsole and exemplar were removed. This was accomplished in a rather rudimentary or primitive way, using the aforementioned graphical user interface, wherein the analyst simply traced the perimeter of the outsole using the cross-hair of the cursor, thus automatically generating a binary image that labeled every pixel as either belonging to the outsole or belonging to the background. Once generated, this map was saved and mathematically multiplied with other images downstream (*e.g.*, the outsole and Handiprint exemplar) to effectively increase image signal to noise ratios. As such, the background (or non-tread areas) of both the outsole and exemplar were removed (Fig. 2) to ensure the highest quality imagery moving forward (*e.g.*, removal of remnants of the analyst's hands that may have been captured during scanning when pressure was applied to the outsole to promote a nearly planar surface, and/or removal of extraneous dust and fingerprints on Handiprint exemplars).

Finally, the outsole and exemplar were collectively translated and rotated to ensure that both were centered within the image frame (8961 × 8961 pixels) and oriented such that the long-axis of the shoe (toe-to-heel) was North-South. This was most easily accomplished using the binary image that was created in the previous step, wherein each pixel was defined as either outsole or background. From this image, the midpoint of the outsole was mathematically computed (x_o, y_o), defined as the x-pixel halfway between the maximum width of the shoe and the y-pixel halfway between the maximum length of the shoe. Since the image frame was 8961 × 8961 pixels, the image frame center was located at pixel coordinate (4481, 4481), so the outsole and Handiprint exemplar images were centered by translating the imagery such that (x_o, y_o) was coincident with (4481, 4481).

To ensure that the shoe's long-axis was North-South, the binary map defining outsole versus background was treated as a bivariate normal distribution, amenable to eigen-decomposition. After decomposition, the resulting eigen-vectors defined the major and minor axes of the best-fit ellipse conforming to the (x, y) coordinates of the pixels that defined the outsole. Ergo, the deviation of the major axes from vertical defined the degree of rotation necessary to ensure that the final imagery was oriented as close to North-South as possible within the image frame.

Following registration and background subtraction, randomly acquired characteristics present on both the outsole and exemplar were marked. This process required the analyst to physically examine each outsole with oblique illumination and 4X magnification. Upon identifying a RAC that appeared on both the outsole and the exemplar, the analyst blacked out the RAC pixels on the Handiprint image using the pencil tool in Adobe® Photoshop® Elements 10. This was completed by tracing the edge of the RAC with the pencil tool (set at 2-pixels wide) and then filling in the RAC (if necessary), with the paint bucket tool while viewing the exemplar at a minimum magnification of 200X. When complete, each feature was examined to ensure that every pixel included within the traced perimeter of the RAC was fully labeled (converted to black). For features found on the edge of the shoe, a lug, or a tread element, the boundary of the RAC was interpolated by hand if the distance for interpolation was short and relatively linear (Fig. 3). In instances when the edge could not be dependably interpolated (*e.g.*, along an irregular segment, a curved surface, or near a large void area), the RAC was traced, but not closed, in order to avoid the introduction of interpolation variability (Fig. 4).

Aside: Thus far, a total of seven analysts have contributed to the generation of this database. Each analyst has earned a minimum of a

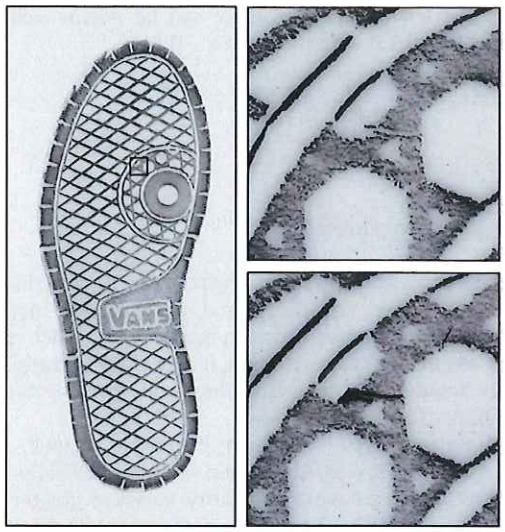


Fig. 3. Illustration of RAC on edge of linear tread element. Note that the edge of the RAC (terminating on the edge of a short and linear tread element), has been interpolated and the entire RAC has been filled in. Vans® sneaker, Skink Mid model, men's size 9.5.

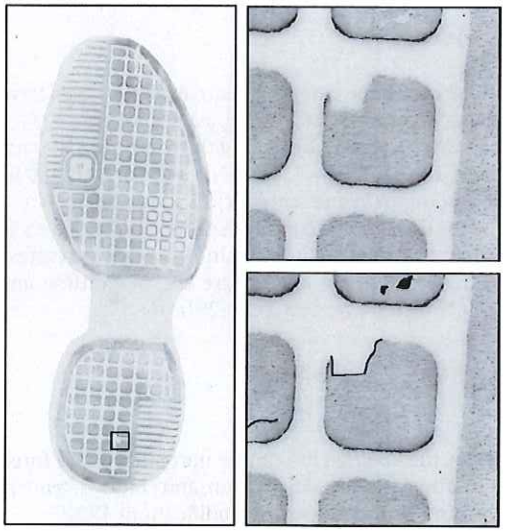


Fig. 4. Illustration of RAC on edge of curved tread element. Note that the edge of the RAC (terminating on the edge of a curved tread element), has not been interpolated nor filled in. Adidas® sneaker, Pro Feather model, men's size 9.0.

Bachelor of Science in Forensic Science from a FEPAC accredited university (two have a Master of Science in Forensic Science from a FEPAC accredited university). All analysts have received 4–16 h of in-house laboratory training by a certified IAI examiner, and 8–16 h of training by a research assistant. However, none are certified examiners and none have completed a full course of study in footwear analysis as recommended by SWGTREAD [24]. The authors acknowledge this as a shortcoming of the research, but without sequestering certified examiners to perform this work (at the expense of casework), a compromise in training was accepted, wherein the research analysts received training devoted almost exclusively to the identification of RAC and subclass characteristics.

When this registered and marked image was subtracted from its registered (but unmarked) counterpart, the result was a RAC map that highlighted the location and geometry associated with each randomly acquired feature (Figs. 5 and 6). Using the standard image processing technique of connected components, the location of each RAC was sequentially characterized using three parameters that were readily available based on x, y pixel coordinates; the radius (r) or distance (in pixels) between the shoe's midpoint and the RAC's centroid (geometric average of the RAC's x, y pixel coordinates), the angular (θ) position (in degrees) between the RAC's centroid and zero degrees (defined as a horizontal line drawn directly East of the shoe's midpoint), and the normalized distance (r_{norm}) equal to r divided by the distance (in pixels) between the shoe's midpoint and the perimeter of the shoe at angular position θ (obtained by casting out a vector from the shoe's midpoint to the shoe's perimeter at angle θ).

Following localization, each feature was automatically numbered (via its connected component value) and extracted from the total RAC map. The resulting subimages (Fig. 7) were then evaluated to define RAC shape and geometry, based on a five-dimensional RAC feature vector, before transformation into individual RAC Fourier descriptors (FD).

2.1. RAC feature vector

Each randomly acquired characteristic was attributed to one of four categories: lines/curves, circles, triangles, and irregular-shaped features. To determine this categorization, five attributes per RAC were required, including area, perimeter, linearity, circularity, and triangularity. The first two descriptions (area and perimeter) were readily available; area describes the total number of pixels comprising the RAC and perimeter evaluates the distance in pixels along a line/curve, or around a two-dimensional shape.

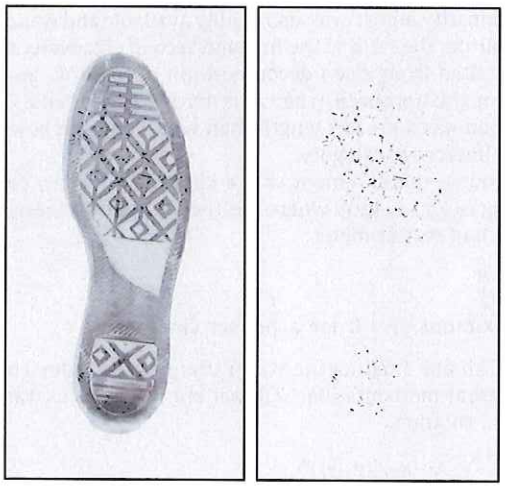


Fig. 5. Registered and marked Handiprint image (left) and resulting RAC map (right).

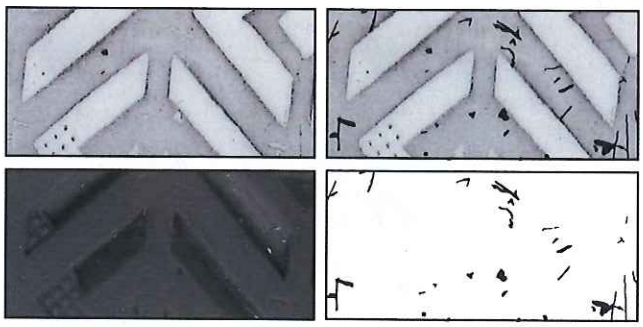


Fig. 6. Example of a selected portion of the Converse Chuck Taylor® All Star®. Handiprint (top left), outsole (bottom left), marked Handiprint (top right), RAC map (bottom right). Note that the outsole image shown in this figure has been scanned on a flat bed scanner, but that all RACs were detected using 4X magnification and oblique illumination.

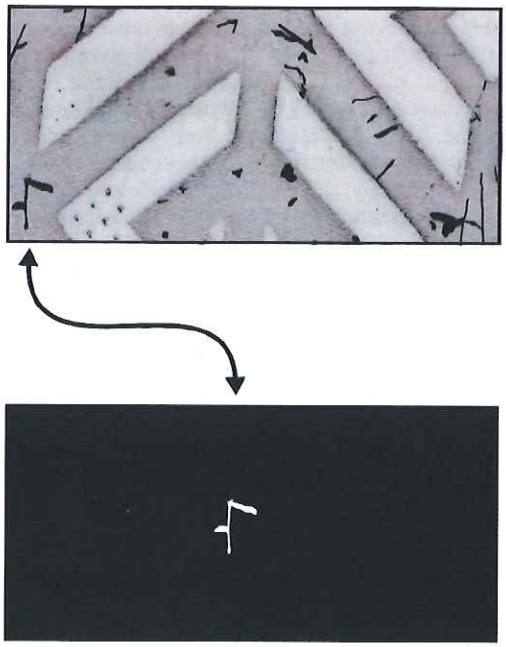


Fig. 7. Subsection of RAC map and example of connected component subimages. This particular RAC was numbered #101, located at a normalized radius of 0.55 and an angle of 104°.

The linearity metric was also readily available and was obtained by computing the ratio of the first and second eigenvalues (λ_1 and λ_2) generated from eigen decomposition of the RAC coordinates [25]. Using this approach, when λ_1 is much greater than λ_2, the RAC in question has a greater length than width and can be classified into the line/curve category.

The fourth measurement was a circularity metric, computed according to Eq. (1) [26], where A is the area of the object, and P is the length of its perimeter:

$$R_c = \frac{4\pi A}{P^2} \qquad (1)$$
$$R_c = \text{maximum of 1.0 for a perfect circle}$$

The fifth and final metric was a triangularity value computed using central moments (Eq. (2)) that are invariant to translation, scale, and rotation.

$$\mu_{pq} = \sum_x \sum_y (x - x_c)^p (y - y_c)^q \qquad (2)$$

As per Rosin [27], the variable I_1 in Eq. (3) equals 1/108 for any triangle that has been affine transformed into a perfect right-angled triangle:

$$I_1 = \frac{\mu_{20}\mu_{02} - \mu_{11}^2}{\mu_{00}^4} \qquad (3)$$

As such, the triangularity measure can be normalized to vary between 0.0 and 1.0 according to Eq. (4) [27]:

$$T = \begin{cases} 108\, I_1 & \text{if } I_1 \le \frac{1}{108} \\ \frac{1}{108\, I_1} & \text{otherwise} \end{cases} \qquad (4)$$

2.2. Categorization parameters

The five-dimensional feature vector (Fig. 8) describing area, perimeter, linearity, circularity, and triangularity served as a primary descriptor and comparison parameter for each randomly acquired characteristic. In addition, it was used to categorize the randomly acquired characteristics into one of four groups; line/curve, circle, triangle, or irregular.

Based on a survey of known geometric shapes, absolute categorization rules were developed. More specifically (and for this dataset), circles have a circularity measure greater than or equal to 0.8, triangles have a circularity measure less than 0.8 *and* a triangularity greater than or equal to 0.9, while lines/curves have a linearity ratio greater than 5 *and* a triangularity measure less than or equal to 0.3; any shape not satisfying one of the above rules defaults into the irregular category (Fig. 9).

2.3. Shape descriptor

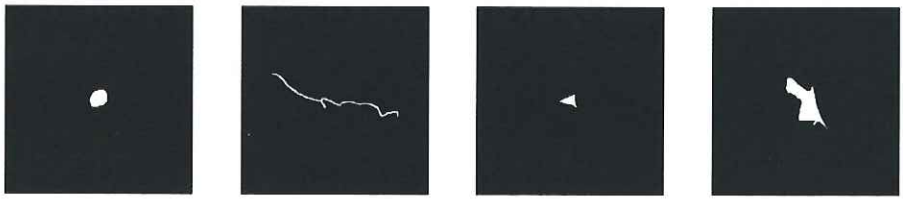
In addition to shape categorization, each RAC was treated as a closed planar figure yielding a Fourier description [28–30]. This description was generated by tracing the contour of the shape ($x(t)$, $y(t)$) (where $t = 0, \ldots, N-1$ with $N = 350$ for this dataset) and assuming a complex plane $z(t) = x(t) + i\,y(t)$ (where $i = \sqrt{-1}$). The resulting one-dimensional complex sequence of numbers was then mapped to the frequency domain via the discrete Fourier transform [29] where R_m and θ_m are the magnitude and phase of the *mth* coefficient, respectively [29]:

$$Z(m) = \sum_{t=0}^{N-1} z(t)\, e^{(-i2\pi mt/N)} = R_m\, e^{(i\theta_m)} \qquad (5)$$
$$m = -N/2, \ldots, -1, 0, 1, \ldots, N/2-1$$

As necessary, the coefficients can be normalized and forced to be invariant to translation, scale, rotation, and contour/sequence start point according to the following modifications [29]:

$$\begin{aligned} &Z(0) = 0 \Rightarrow \text{translation invariance} \\ &R_m = \frac{R_m}{R_1} \Rightarrow \text{scale invariance} \\ &\theta_m = \theta_m - \frac{\theta_{-1} + \theta_1}{2} \Rightarrow \text{rotation invariance} \\ &\theta_m = \theta_m + m\frac{\theta_{-1} - \theta_1}{2} \Rightarrow \text{start point invariance} \end{aligned} \qquad (6)$$

To illustrate, consider Figs. 10 and 11. Fig. 10 depicts a single RAC (A), along with four synthetic modifications (B–E showing changes in scale, rotation, and translation). The resulting normalized



[299, 63, 0.93, 0.72, 1.54] Circle
[395, 145, 0.04, 0.01, 77.99] Line/Curve
[132, 54, 0.56, 0.95, 2.31] Triangle
[917, 188, 0.33, 0.47, 6.97] Irregular

Fig. 8. Four RAC images with their corresponding feature vectors (area, perimeter, circularity, triangularity, linearity).

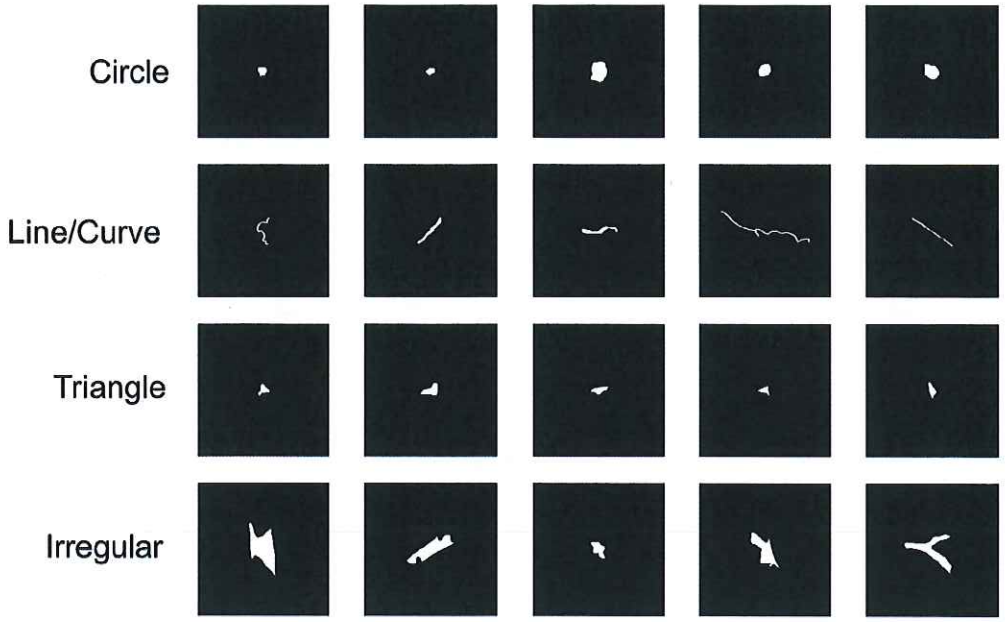


Fig. 9. Examples of RACs classified as circles, lines/curves, triangles, and irregulars.

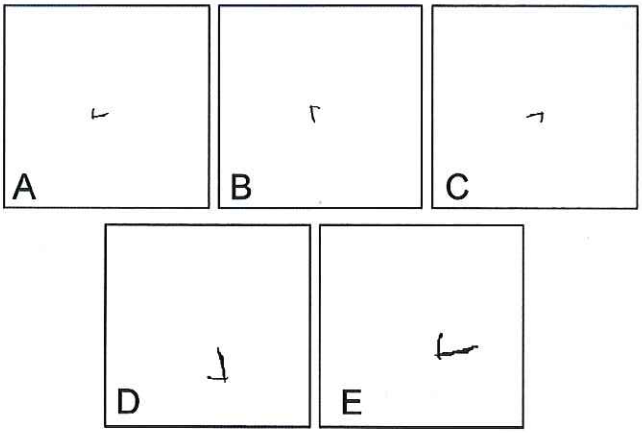


Fig. 10. (A) Original RAC, (B) rotated, (C) rotated, (D) rotated, translated, and scaled, (E) scaled and translated.

Fourier descriptors are plotted in Fig. 11. The x- and y-axes are arbitrary dimensions since the images have been normalized, but note that all contours are normalized to the same configuration, save a single π radian ambiguity [31]. Unless otherwise noted, all subsequent uses of RAC Fourier descriptors make use of both translation and start point invariance modifications.

3. Frequency and similarity assessment

At this point in data acquisition, each RAC has a geometric description, location and well-defined origin (from a left or right shoe with a known pattern, a known manufacturer (usually), a known size, etc.) and can be assessed as such. However, for any given shoe size (or pattern, or brand, etc.) the database itself is limited in sample size. With this in mind, an interim solution (at least until the database grows to such a size that sampling is considered robust) is to transform the frequency information into a normalized space that allows for numerical assessment regardless of shoe size, shape, pattern, etc. Naturally, this simplification bounds the utility of the frequency information, and the authors urge the user to be cognizant of this moving forward, but the transformation in no way invalidates provisional usefulness.

3.1. Outsole size and shape normalization

Normalization was achieved using a *single idealized shoe* corresponding to a men's size 10 Reebok® walking shoe with an outsole surface area of 21,235 mm^2. Beginning from the top medial portion of the shoe, the outsole was divided into 5 mm × 5 mm cells through a rastering process, creating 990 total cells of which 860 were complete, and 130 were partial (or straddling the perimeter/edge of the outsole as illustrated in Fig. 12). By mapping between Cartesian and polar coordinates, each RAC could be localized via θ and r_{norm}. Essentially, this meant that a RAC near the edge of the medial part of the heel on a women's size 6.5 could have the same θ and r_{norm} as a RAC on the edge of the medial part of the heel of a men's size 10.0, and therefore map to the same 5 mm × 5 mm cell in the normalized outsole. (*Note: we also have the capacity to report frequency values as absolute, physical or non-normalized values using θ and r. This would be equivalent to taking a stack of Handiprints, centering all shoes in the middle of each sheet with the toe-heel oriented North-South, and drilling down through all sheets at a fixed location, regardless of shoe size. To further elaborate, in the aforementioned example, the RAC on the medial heel portion of the women's size 6.5 shoe would likely fall somewhere in the lower-instep area of the men's size 10.0.)*

3.2. Similarity assessment

The aforementioned normalization step yields RAC frequency information and the potential for chance co-occurrence of RACs within a 5 mm × 5 mm cell on an outsole (in other words, the dataset can empirically estimate the random chance of discovering two or more accidentals in the same position on shoes previously known to be unrelated). This can be further divided by geometry in terms of the chance co-occurrence of lines/curves, circles, triangles, or irregular shaped RACs within a 5 mm × 5 mm cell. However, chance co-occurrence in position and general category (line/curve, circle, triangle, or irregular shape) does not mean coincidental association in actual geometry since general categorization does

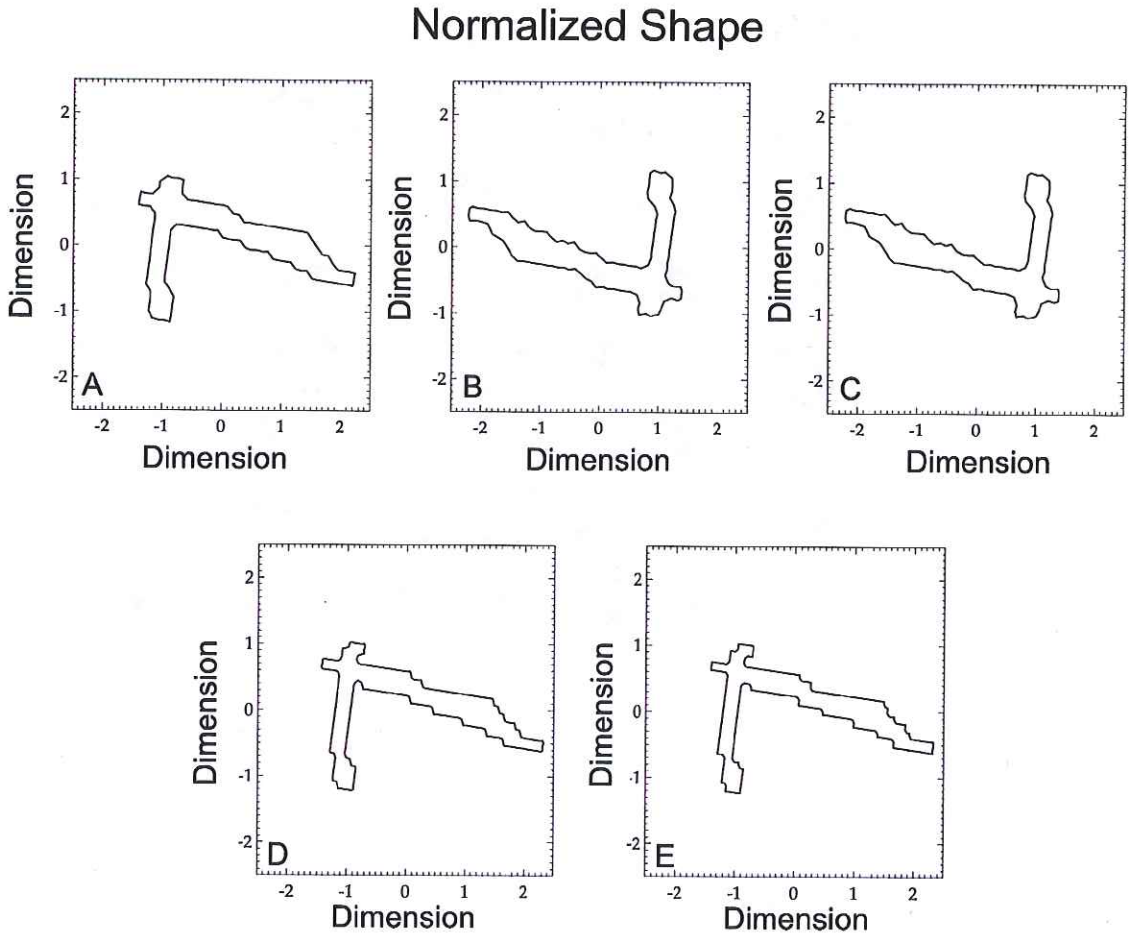


Fig. 11. Plot of normalized Fourier shapes derived from the RACs shown in Fig. 10.

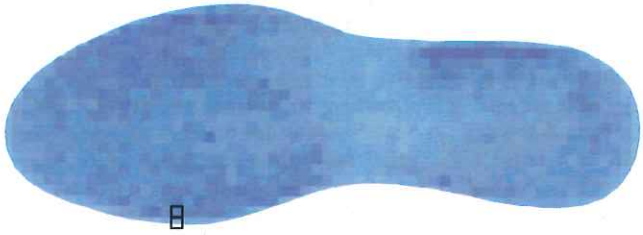


Fig. 12. Illustration of a single full and partial 5 mm × 5 mm cell on the normalized outsole.

not sufficiently describe RAC complexity. In order to further assess the similarity in shape for RACs that co-exist within a given cell by chance, a numerical metric of similarity can be utilized to pre-rank the data for the analyst, thereby efficiently limiting the number of visual pair-wise comparisons required. To accomplish this, modified phase only correlation (MPOC) was employed. This metric utilizes the Fourier transform $F[g(x,y)] = G(u,v)$ of the RAC spatial domain image $g(x, y)$ giving the analyst access to frequency information associated with the RAC's amplitude $A(u, v)$ and phase $\sigma(u, v)$ as illustrated in Eq. (7) where $i = \sqrt{-1}$ [10].

$$G(u,v) = A(u,v)e^{i\sigma(u,v)} \tag{7}$$

With this in mind, the similarity ($POC_{g_1g_2}$) between two RAC images – $g_1(x, y)$ and $g_2(x, y)$ – can be determined (a value between 0.0 and 1.0) according to Eq. (8) [4,7,9] where F^{-1} is the inverse Fourier transform and G_2^* is the complex conjugate of G_2 [10].

$$POC_{g_1g_2} = F^{-1}\left[\frac{G_1(u,v)G_2^*(u,v)}{|G_1(u,v)G_2^*(u,v)|}\right] = F^{-1}[e^{i[\sigma(u,v)-\theta(u,v)]}] \tag{8}$$

As a "tuning" step, Eq. (8) can be modified by application of a frequency filter that selectively limits frequencies used in the computation such that $F[g(x,y)\cdot h(k,l)] = G(u,v)$. In this instance, each image $g(x, y)$ is modified by the windowing function shown in Eq. (9) with $\alpha = 0.2$ and where $k = l = N$ which is the size of the RAC spatial domain image in pixels (1600 × 1600):

$$h(k) = \alpha - (1-\alpha)\cos\left[\frac{2\pi k}{N}\right] \quad k = 0, 1, \ldots, N-1 \tag{9}$$

4. Results

4.1. Database statistics

To date, more than 1000 shoes have been pre-processed. The defining characteristics of the first 1000 (501 lefts and 499 rights) are detailed in Tables 1–6. The majority of shoes in this collection are athletic in nature (Table 1), due to generous corporate donations and the availability of shoes for purchase from local thrift stores. Table 2 reports the degree of wear of each shoe, which is not quite balanced between light, moderate, and heavy. For this study, shoes with "light wear" are those that exhibit discernible texture throughout. Conversely, the label "moderate wear"

describes shoes with a reasonable degree of wear, resulting in both lost texture and possible bald spots. Finally, the term "heavy wear" is reserved for shoes with a near complete loss of texture, many or large bald spots, and possible holes or areas where the outsole has completely worn through.

Table 3 shows that nearly 90% of the collection lacks microcellular material in outsole composition. This is fortuitous since the presence of microcellular material is likely to increase intra- and inter-analyst variability in identifying randomly acquired characteristics. Conversely, approximately three-quarters of the database show Schallamach patterns (Table 4); this is likewise fortuitous. Although current RAC data does not include the quantification of these features, the discrimination potential of Schallamach patterns can be explored in future studies.

Table 5 reports shoe frequency as a function of manufacturer and/or brand. Results indicate that almost 30% of the shoes processed thus far are from Nike® while another 28% are comprised of a small number of shoes, but from numerous manufacturers. Finally, Table 6 breaks down the database according to size and intended market (men or women). The results here are not random, but selective in the sense that our group did not capture data for shoes with a physical outsole size greater than the maximum length of a sheet of Handiprint currently available for purchase (or approximately 13 inches in total length).

The shoes in Table 1 generated a total of 57,426 RACs (average of 57, minimum of 1, and maximum of 410). The majority (45%) were categorized as lines/curves, with another 38% falling into the irregular category. Circles filled a distant third group, comprising only 11% of the database, with triangles completing the remaining 6% (Table 7).

The agreement between "automated" and "human" categorization of RACs ranged between 95% and 68%, depending on the complexity and imperfections of the shape under review. For example, using a test set of 74 "stylized" shapes (manually created in ImageJ [32] with an intended geometry), plus 110 randomly selected RACs, the overall agreement or accuracy in categorization was computed to be 95%. This was determined by taking the total test set of 184 images and presenting them via a graphical user interface to analysts seated at a computer. When presented with each image, in a randomized order, the analyst was asked to categorize the shape as either a circle, triangle, line/curve or irregular-shaped feature by clicking on a corresponding toggle button. The same shape was automatically categorized using the automated decision rules determined during our training phase, and the results for three separate analysts (for a total of 552 human-perceptual estimates of shape categorization) were combined into the confusion matrix shown in Table 8 with an overall agreement of 95%.

Conversely, for a total of 800 randomly selected RACs (zero stylized shapes), assessed by four analysts (200 each, with a total of 746 human-perceptual estimates of shape categorization of which 27 RACs happen to repeat during the randomized selection), the equivalent confusion matrix (shown in Table 9) was found to have an overall agreement of 68%. Despite the clear decrease in agreement, the authors assert that this should not be defined as an "error rate" since it is based on human-perception of shapes, which cannot be expected to agree among or between individuals. The problem is that there is no appropriate reference by which to define "ground truth" as soon as shapes become complex and imperfect. To illustrate this, consider Figs. 13 and 14 which show a sampling of RAC images that lead to disagreement in the human-perception versus automated algorithm study, contributing to the results shown in Table 9. In both figures, the top row denotes the automated categorization label, while the cell label indicates the human analyst choice. Depending on the viewer and the image, there are some instances where the human's reasoning seems more "accurate," and some instances where the algorithm's choice seems more "accurate." Overall, the results suggest that it would not be robust to keep a large number of RAC shape groupings, due to both human-perceptual differences and RAC complexity/imperfections. Given this observation, the authors suggest a maximum of three groups that may be useful moving forward; "irregular" for complex structures, "elongated" to describe lines and curves, and a new grouping defined as "approximate isometry" to include circular and triangular structures.

Table 7
Frequency of RACs by shape category.

Metric	All RACs	Irregulars	Circles	Triangles	Lines/curves
Total	57,426	22,075	6287	3242	25,822
Percentage	100%	38%	11%	6%	45%
Minimum number in a cell	0	0	0	0	0
Maximum number in a cell	132	52	21	14	78
Mean number in a cell	58	22	6	3	26
Median number in a cell	61	23	6	3	27

Table 8
Confusion matrix for automated categorization of 184 shapes (74 stylized and 110 real RACs) as assessed by three analysts for a total of 552 human-perceptual assessments of shape. The column headers represent the algorithm report while the rows designate human-perception. Total agreement equals 95%.

	Circle	Triangle	Line/curve	Irregular
Circle	99	0	0	0
Triangle	0	90	2	4
Line/curve	0	0	214	8
Irregular	1	5	5	124

To assess inter- and intra-analyst variation in RAC marking, a random set of 100 pairs of shoes (approximately 10% of the database) were selected for periodic reassessment. On an approximate bimonthly basis, each analyst selected the next available shoe from the randomized list (which may or may not be a shoe he or she has already marked), and repeated the marking process on the post-registered and background subtracted image. Subtraction of the newly marked RAC image from its registered and unmarked mate created a secondary RAC map. Differences between replicate maps then served as a basis for assessing inter- and intra-analyst variation in marking.

Thus far, the quality assessment program has obtained 160 paired RAC maps, prepared by five analysts, over a 15-month time period. The information contained in each of the 320 RAC maps (two markings × 160 shoes) has been assessed in two ways. First, the data has been converted into a one-dimensional (1D) vector by rastering across image rows and down image columns, collecting total RAC size per cell using a fixed bin width of 150 × 150 pixels (approximately 6 mm × 6 mm). The resulting 1D feature vectors of RAC size (per cell) for paired RAC maps were then evaluated to determine the average correlation coefficient of similarity. Inter-analyst variation produced an average correlation coefficient of 0.66 with a variance of 0.057, based on 137 paired RAC maps. To date, the dataset has allowed for the computation of intra-analyst correlation, but thus far, based on only 23 paired RAC maps for two analysts in the research group; the combined average correlation coefficient is 0.80 with a variance of 0.016. In addition to the image-wide correlation scores, individual uncertainty of measure for θ, r, and r_{norm} has been computed based on this same dataset. Table 10 reports the mean, variance and range of

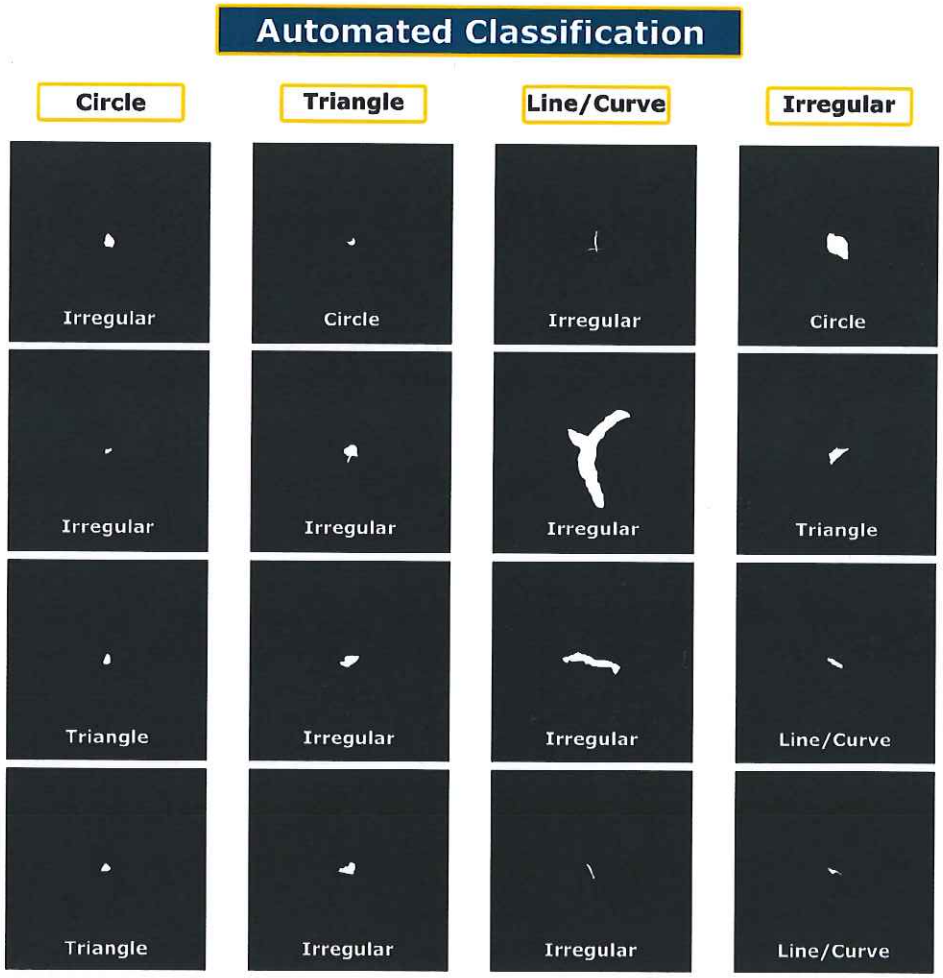


Fig. 13. Illustration of disagreement in human-perception of shape categorization (cell labels) versus automated categorization based on training rules (column header). To account for this disagreement a reduction in grouping complexity (from four to three) is suggested: irregular, elongated (lines and curves) and approximate isometry (a combination of circular and triangular structures).

measurements associated with each value based on 160 shoes and 5477 duplicate marked randomly acquired characteristics (combined inter- and intra-analyst markings). In addition, illustrations of duplicate markings of known match RACs are shown in Fig. 15, along with individual measurement differences. The results indicate that angular differences are very small (less than a 1°) and that radial distances differ by 0.16 ± 1.9 mm. The interpretation of each quality metric (correlation versus measurement uncertainty) indicates that the greatest variation is within the RAC *detection* process; however, when a RAC is detected, on average, it is *consistently marked* in the same manner by all analysts in the research group.

Table 11 details RAC frequency in each of the 990 bins on the normalized shoe; approximately 2.4% of the cells are empty (or void of RACs), and approximately 4.5% of the cells contain five or fewer features, leading to at most a 1 in 11,485 random chance of RAC co-occurrence in position (without regard for shape). Moreover, the majority of bins (80%) contain between 26 and 100 RACs with a mean of approximately 58 features per cell (Table 7). The results of the quality assurance program indicate that in-house training ensures that analysts mark RACs in a highly consistent manner, but that we have less control over the standardization of day-to-day RAC *detection* (which is a function of lighting, magnification, fluctuations in analyst attention, fatigue, etc.). Given this, the next uncertainty estimate to be explored in a companion paper is actual frequency or count (*e.g.*, the uncertainty in the frequency of chance co-occurrence or 1:1000 ± *X*).

In addition to the summary statistics provided in Tables 7 and 11, RAC frequency was also globally localized as a function of shoe section as illustrated in Fig. 16. More specifically, the normalized shoe was divided into eight sections, equally bisecting the shoe into medial and lateral sections, as well as four quarters from heel to toe. Tables 12–19 report the total number of acquired features per section. Overall, there were more RACs within the toe (36,346 features) than the heel (21,080 features). In addition, the most populated area was section 3, or the lower, lateral toe area of the outsole, which contained approximately 18% of all RACs (Table 14) and a mean RAC frequency of 69 per 5 mm × 5 mm cell. Conversely, the least populated area was section 6, or the medial arch/upper heel region of the outsole, which matches intuition since depending on shoe design, this region can have minimal contact with the ground, and thus, a lower potential to acquire features. Within this area, there was an average of only 34 features per 5 mm × 5 mm cell, with a total of 3886 RACs (Table 17).

For this database, the *greatest potential for random chance co-occurrence in RAC position and shape category* was found to exist for a single 5 mm × 5 mm bin located in section 2, approximately 70 mm from the heel of the normalized shoe. This particular area on the outsole had a probability of co-occurrence that ranged from 1:756 to 1:9571 as illustrated in Table 20. Given this potential for

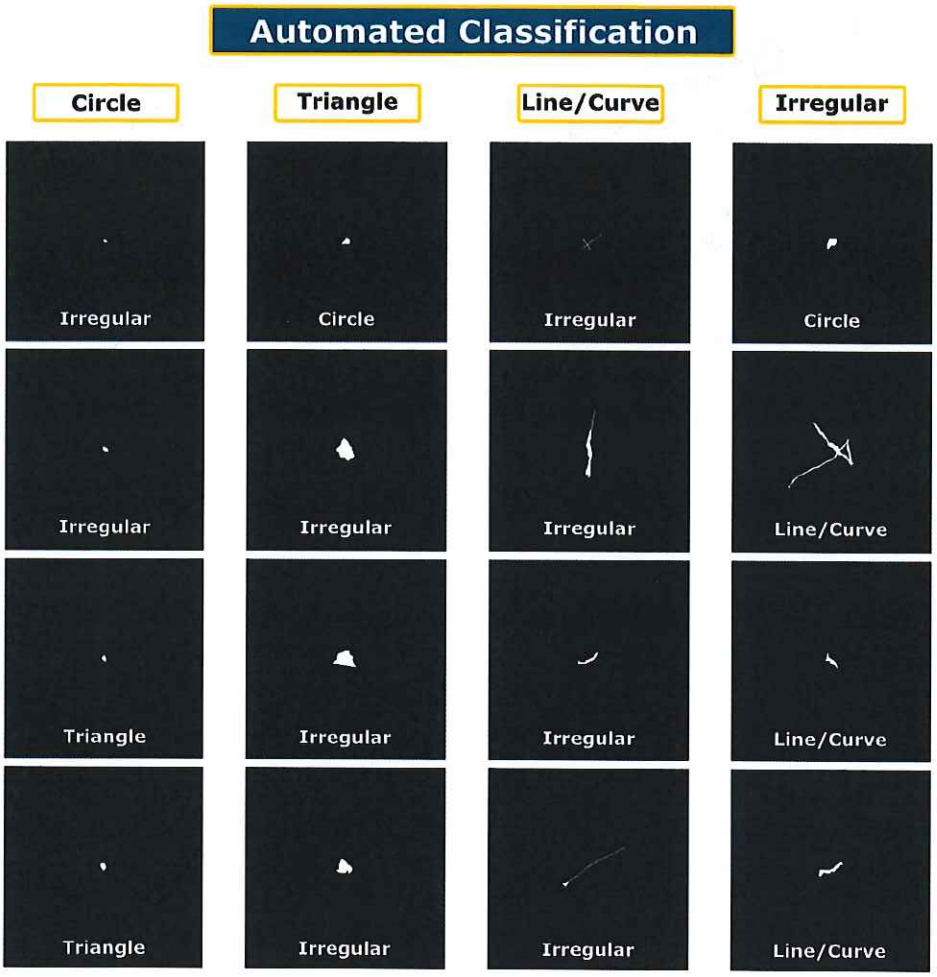


Fig. 14. Illustration of disagreement in human-perception of shape categorization (cell labels) versus automated categorization based on training rules (column header). To account for this disagreement a reduction in grouping complexity (from four to three) is suggested: irregular, elongated (lines and curves) and approximate isometry (a combination of circular and triangular structures).

co-occurrence, all RACs localized to this bin were pairwise compared and ranked in terms of similarity using modified phase only correlation (MPOC). For convenience, the *two most similar RACs* were visually reproduced for the analyst in Fig. 17 (meaning all other pairwise comparisons had geometries less similar than those shown). The visual representation provides the viewer with the reported MPOC scores, the actual RAC images, as well as the associated Fourier images. Although *some level of visual similarity* can be discerned, the accidental features are distinguishable based on size, shape and/or orientation. However, the fact that *some level of expressed similarity* is apparent should not be ignored, and clearly much more work is required to better understand the limit of discrimination as a function of RAC size and complexity following positional chance association.

Given that there are 990 cells on the normalized shoe, a total of 57,426 RACs, and a resulting 2,021,440 paired-similarity comparisons, the volume of information that can be gleaned from this dataset is enormous and requires some type of user-friendly interface. Toward this end, an interactive web-based heat map has been created, of which a static-beta version is illustrated in Figs. 17 and 18 (see the following URL for the most up-to-date link www.4n6chemometrics.com/database/). The goal is to report RAC co-occurrence per cell, per shape category, as well as MPOC scores and associated imagery (actual and normalized Fourier). This in turn allows the analyst to visually and quantitatively evaluate the spatial density of randomly acquired characteristics according to location and shape in response to the National Academy of Sciences' (NAS) 2009 request for relative frequency of features, as well as SWGTREAD's request for research on "*Random Placement Shape and/or Placement of Randomly Acquired Characteristics*" [22].

Table 9
Confusion matrix for automated categorization of 746 unique RACs and 27 repeated RACs as assessed by four analysts for a total of 800 human-perceptual assessments of shape. The column headers represent the algorithm report while the rows designate human-perception. Total agreement equals 68%. Note that "rectangle" was an early subdivision of the line/curve category, that was later phased out (i.e., rectangles = line/curve).

	Circle	Triangle	Rectangle	Line/curve	Irregular
Circle	46	10	0	0	19
Triangle	12	25	0	1	19
Rectangle	3	3	0	2	3
Line/curve	0	7	0	340	58
Irregular	9	80	0	29	134

Table 10
Variation in analyst duplicate marking of 5477 randomly acquired characteristics across 160 shoes (320 RAC maps).

Metric	θ (degrees)	r (pixels)	r (mm)	r_{norm}
Mean	0.0922	4.27	0.167	0.00177
Variance	0.0178	91.7	3.61	0.0000121
Maximum	0.699	112	4.40	0.0300

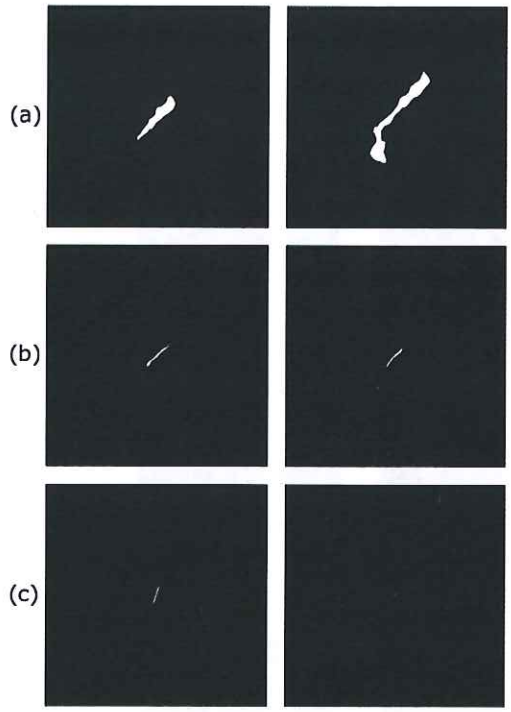


Fig. 15. Duplicate markings of known match RACs with the following marking variations: (a) $\Delta\theta = 0.235$, $\Delta r = 67.2$ px/2.64 mm, $\Delta r_{norm} = 0.0170$; (b) $\Delta\theta = 0.567$, $\Delta r = 88.5$ px/3.49 mm, $\Delta r_{norm} = 0.0270$; (c) $\Delta\theta = 0.551$, $\Delta r = 112$ px/4.40 mm, $\Delta r_{norm} = 0.0290$.

Table 11
Frequency of RAC counts in 5 mm × 5 mm bins across a normalized shoe containing 990 total bins which are at least partially located on the outsole. In addition, the potential for random duplication of RAC position, based on this database, is provided.

Number of RACs (potential for co-occurrence)	Frequency	Percent of total
0 (1:–)	24	2.4%
1 (1:57,426)	4	0.4%
2–5 (1:28,713–1:11,485)	18	1.8%
6–10 (1:9571–1:5742)	17	1.7%
11–25 (1:5220–1:2297)	91	9.2%
26–50 (1:2208–1:1148)	196	19.8%
51–75 (1:1126–1:765)	356	36.0%
76–100 (1:568–1:574)	242	24.4%
101–132 (1:568–1:382)	42	4.2%
Total	990	100%

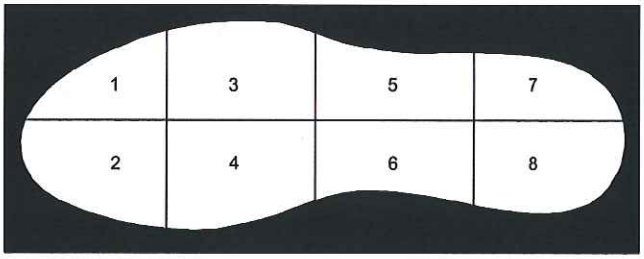


Fig. 16. An illustration of the normalized shoe outsole broken into eight sections. The horizontal line equally bisects the shoe, while the vertical lines divide the shoe into quarters.

and the "*Mathematical Probabilities of Randomly Acquired Characteristics*" [23].

However, the authors acknowledge that the database must be used with caution. The utility of the density information is its

Table 12
Frequency of RACs by shape category in section 1, as illustrated in Fig. 16.

Metric	All RACs	Irregulars	Circles	Triangles	Lines/curves
Total	7309	2837	755	428	3289
Minimum number in a cell	0	0	0	0	0
Maximum number in a cell	116	51	19	14	64
Mean number in a cell	70	27	7	4	32
Median number in a cell	82	32	8	4	35

Table 13
Frequency of RACs by shape category in section 2, as illustrated in Fig. 16.

Metric	All RACs	Irregulars	Circles	Triangles	Lines/curves
Total	8477	3228	849	495	3905
Minimum number in a cell	0	0	0	0	0
Maximum number in a cell	113	49	21	10	68
Mean number in a cell	61	23	6	4	28
Median number in a cell	69	25	6	3	31

Table 14
Frequency of RACs by shape category in section 3, as illustrated in Fig. 16.

Metric	All RACs	Irregulars	Circles	Triangles	Lines/curves
Total	10,377	4002	1170	568	4637
Minimum number in a cell	0	0	0	0	0
Maximum number in a cell	124	46	19	14	72
Mean number in a cell	69	27	8	4	31
Median number in a cell	68	27	8	4	31

Table 15
Frequency of RACs by shape category in section 4, as illustrated in Fig. 16.

Metric	All RACs	Irregulars	Circles	Triangles	Lines/curves
Total	10,183	3956	1121	574	4532
Minimum number in a cell	0	0	0	0	0
Maximum number in a cell	113	48	17	12	64
Mean number in a cell	67	26	7	4	30
Median number in a cell	69	27	7	4	28

Table 16
Frequency of RACs by shape category in section 5, as illustrated in Fig. 16.

Metric	All RACs	Irregulars	Circles	Triangles	Lines/curves
Total	4935	1886	491	264	2294
Minimum number in a cell	0	0	0	0	0
Maximum number in a cell	132	52	18	8	76
Mean number in a cell	44	17	4	2	20
Median number in a cell	32	15	4	2	14

Table 17
Frequency of RACs by shape category in section 6, as illustrated in Fig. 16.

Metric	All RACs	Irregulars	Circles	Triangles	Lines/curves
Total	3886	1610	425	184	1667
Minimum number in a cell	0	0	0	0	0
Maximum number in a cell	99	32	17	6	48
Mean number in a cell	34	14	4	2	15
Median number in a cell	33	14	3	1	12

Table 18
Frequency of RACs by shape category in section 7, as illustrated in Fig. 16.

Metric	All RACs	Irregulars	Circles	Triangles	Lines/curves
Total	5414	1995	630	321	2468
Minimum number in a cell	0	0	0	0	0
Maximum number in a cell	131	44	17	11	78
Mean number in a cell	59	22	7	4	27
Median number in a cell	59	24	7	4	26

Table 19
Frequency of RACs by shape category in section 8, as illustrated in Fig. 16.

Metric	All RACs	Irregulars	Circles	Triangles	Lines/curves
Total	6845	2561	846	408	3030
Minimum number in a cell	1	0	0	0	0
Maximum number in a cell	87	36	21	11	49
Mean number in a cell	53	20	7	3	23
Median number in a cell	55	21	6	3	24

Table 20
Frequency of RACs and potential for co-occurrence as a function of position and shape for bin located in section 2, approximately 70 mm from the heel of the shoe.

Description	Any shape	Irregular	Circle	Triangle	Line/curve
Total: in database	57,426	22,075	6287	3242	25,822
Total: in cell	132	39	11	6	76
Chance of finding RAC in cell	1:435	1:1472	1:5220	1:9571	1:756

ability to shed light on the random and variable nature of RAC frequency and *possible co-occurrence*. However, the heat map data is not intended to be a quantitative collection of independent wear-related events that can be multiplied to provide a cumulative probability of occurrence for a constellation of RACs on a randomly

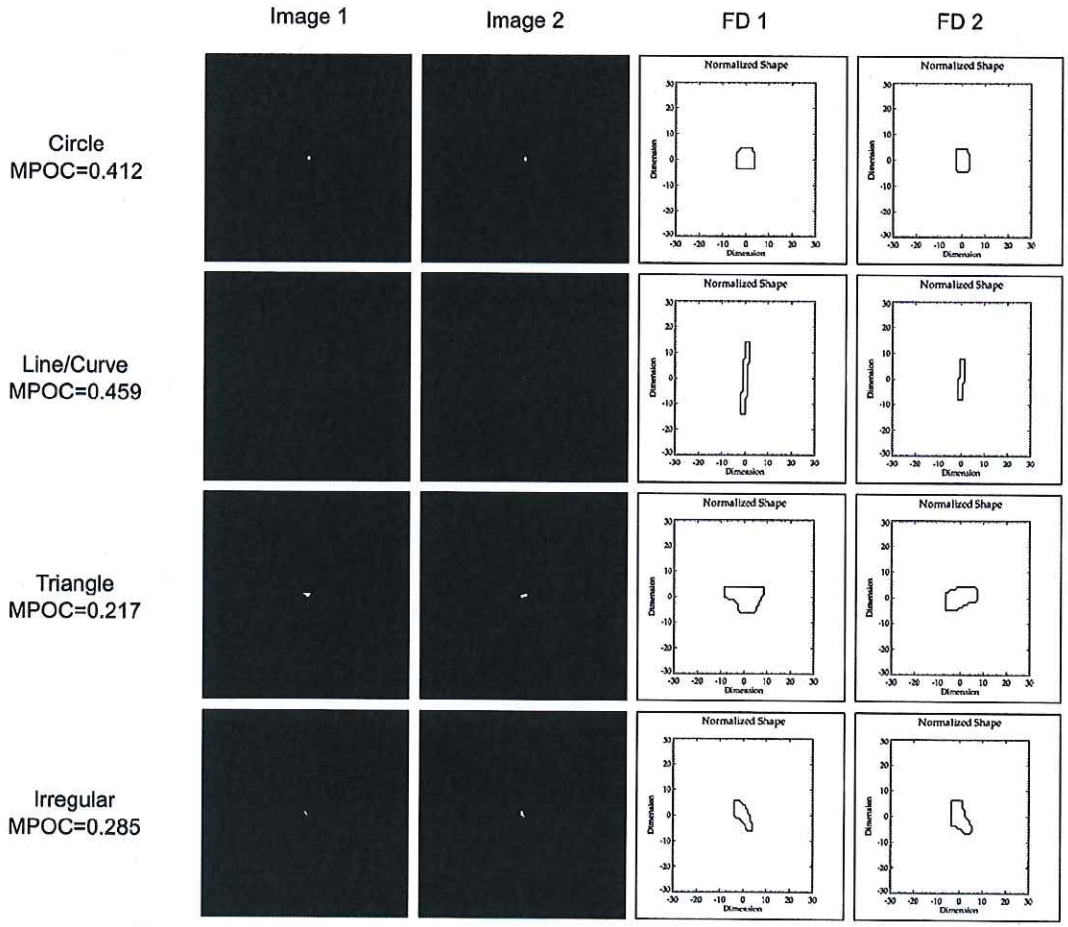


Fig. 17. An illustration of the most similar RACs (*i.e.*, highest MPOC score) in each shape category within the bin located in section 2, approximately 70 mm from the heel of the shoe. The two RAC images are displayed in the first two columns. In addition, the Fourier descriptors (FD) for both images are included for easier visualization (last two columns). Note that the most similar RACs are distinguishable based upon visual inspection and a correspondingly low MPOC score.

Description	Any Shape	Irregular	Circle	Triangle	Line/Curve
Total: In Database	57,426	22,075	6,287	3,242	25,822
Total: In Cell	86	35	12	0	39
Chance of Finding RAC in Cell	1:667	1:1,640	1:4,785	1:-	1:1,472

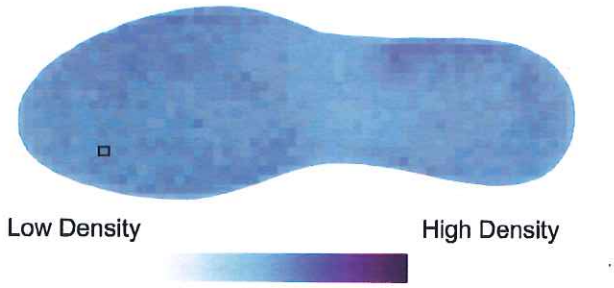


Fig. 18. Static illustration of web-based heat map for a normalized shoe. Numerical values in the top row of the associated frequency table remain constant regardless of the user's interaction with the heat map, displaying data associated with total RAC count for the entire database (regardless of cell location). Conversely, the middle and bottom rows automatically update to display RAC count and frequency for individual cells (5 mm × 5 mm) when queried by the user. In this static example, the results are shown for a single cell outlined in black near the toe. Note that the normalized shoe was a size 10 men's Reebok® walking shoe with an area of 21,235 mm^2. Following cell selection, the user is then able to navigate to a second web-page that illustrates RAC similarity.

selected outsole. Moreover, density and categorization does little to account for the clarity, quality, and complexity of a geometric feature, which is as much (if not more important) to the forensic footwear comparison than the simple assessment of presence or absence. As such, the examiner's responsibilities cannot be deduced to a simple table of frequencies, and a great deal more is required to both interpret and understand how best to utilize the database this project is generating. Despite this caveat, now that the data exists and is accessible to the community, our new focus is how best to present it to maximize value, along with estimates of uncertainty in frequency, continued characterization of analyst-variability, and quantitative metrics of shape similarity. To address these concerns, additional data and research is sought. The *ideal* end goal is a detailed analysis of co-occurrence in position and shape, fully accessible via an online interface similar to that shown in Figs. 17 and 18, along with recommendations regarding limits in discrimination as a function of RAC size, area, geometry, and complexity.

Acknowledgments

This project was supported by Award No. 2013-DN-BX-K043, awarded by the National Institute of Justice (NIJ), Office of Justice Program, U.S. Department of Justice. The opinions, findings, conclusions and recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice. In addition to NIJ, the authors would also like to thank several individuals and companies for providing generous corporate donations of used and returned footwear. In alphabetical order, a sincere thank you to Jamie Bragg (Under Armour Inc.), Phil Gallant (Senior Director of Development Teva & Simple Deckers Outdoor Corporation), Herb Hedges (Nike World Headquarters) and Bob Rich (Director of Research & Engineering for Reebok International). Last but not least, the authors would like to thank Rose Paris, the District Manager of Goodwill for Southwestern Pennsylvania, as well as all local Morgantown, WV Goodwill store associates for their assistance in shoe purchases.

References

[1] W.J. Bodziak, Footwear Impression Evidence: Detection, Recovery and Examination, second ed., CRC Press, 2000.
[2] A. Alexander, A. Bouridane, D. Crookes, Automated classification and recognition of shoeprints, Image Process. Appl. 2 (465) (1999) 638–641.
[3] Z. Geradts, J. Keijzer, The image-database REBEZO for shoeprints with developments on automatic classification of shoe outsole designs, Forensic Sci. Int. 82 (1996) 21–31.
[4] P. de Chazal, J. Flynn, R. Reilly, Automated processing of shoeprint images based on the Fourier transform for use in forensic science, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 341–350.
[5] T. Kiely, Forensic Evidence: Science and the Criminal Law, second ed., CRC Press, 2006.
[6] T. Hannigan, L. Fleury, R. Reilly, B. O'Mullane, P. de Chazal, Survey of 1276 shoeprint impressions and development of an automatic shoeprint pattern matching facility, Sci. Justice 46 (2006) 79–89.
[7] M. Gueham, A. Bouridane, D. Crookes, Automatic recognition of partial shoeprints based on phase-only correlation, IEEE 4 (2007) 441–444.
[8] M. Gueham, A. Bouridane, D. Crookes, O. Nibouche, Automatic recognition of shoeprints using Fourier–Mellin transform, in: NASA/ESA Conference on Adaptive Hardware and Systems, 2008, 487–491.
[9] R. Xiao, P. Shi, Computational Forensics: Lecture Notes in Computer Science, vol. 5158, Springer, 2008.
[10] A. Bouridane, Imaging for Forensics and Security: From Theory to Practice, Springer, 2009.
[11] M. Jing, W. Ho, L. Chen, A novel method for shoeprints (sic) recognition and classification, in: Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, 2009, pp. 2846–2851.
[12] M. Cassidy, Footwear Identification, Lightning Powder Company, Inc., 1995 (Reprint).
[13] T.W. Adair, J. Lemay, A. McDonald, R. Shaw, R. Tewes, The Mount Bierstadt study: an experiment in unique damage formation in footwear, J. Forensic Identif. 57 (2) (2007) 199–205.
[14] N. Petraco, C. Gambino, T. Kubic, D. Olivio, N. Petraco, Statistical discrimination of footwear: a method for the comparison of accidentals on shoe outsoles inspired by facial recognition techniques, J. Forensic Sci. 55 (2010) 34–41.
[15] H. Wilson, Comparison of the individual characteristics in the outsoles of thirty-nine pairs of Adidas Supernova Classic shoes, J. Forensic Identif. 62 (3) (2012) 194–203.
[16] H.D. Sheets, S. Gross, G. Langenburg, P.J. Bush, M.A. Bush, Shape measurement tools in footwear analysis: a statistical investigation of accidental characteristics over time, Forensic Sci. Int. 232 (2013) 84–91.
[17] M. Marvin, A look at close non-matching footwear examinations, Presented at International Association for Identification (IAI) Centennial Conference Sacramento, CA, 2015.
[18] Y. Yekutieli, Y. Shor, S. Wiesner, T. Tsach, Expert Assisting Computerized System for Evaluating the Degree of Certainty in 2D Shoeprints, Technical Report, TP-3211, National Institute of Justice, 2012.
[19] Y. Shor, S. Wiesner, Methodological shift in scientifically evaluating shoeprint: a computerized system to statistical aid the expert reach the degree of certainty, Presented at International Association for Identification (IAI) Centennial Conference Sacramento, CA, 2015.
[20] Federal Bureau of Investigation Footwear (Boot) Database, Personal Communication, 2016.
[21] NAS, Strengthening Forensic Science in the United States: A Path Forward; Committee on Identifying the Needs of the Forensic Sciences Community, Technical Report, National Research Council, 2009, https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf.
[22] SWGTREAD, Scientific Working Group for Shoeprint and Tire Tread Evidence – Recommendations for Research: Random Placement Shape and/or Placement of Randomly Acquired Characteristics. http://www.swgtread.org/research/recommendations-for-research/8-footwear/50-random-placement-shape-and-or-placement-of-randomly-acquired-characteristics (accessed November 2015).
[23] SWGTREAD, Scientific Working Group for Shoeprint and Tire Tread Evidence – Recommendations for Research: Mathematical Probabilities of Randomly Acquired Characteristics. http://www.swgtread.org/research/recommendations-for-research/10-footwear-and-tires/42-mathematical-probabilities-of-randomly-acquired-characteristics (accessed November 2015).
[24] SWGTREAD, Guide for Minimum Qualifications and Training for a Forensic Footwear and/or Tire Tread Examiner, 2006, http://www.swgtread.org/images/documents/standards (accessed February 2016).
[25] U. Park, A.K. Jain, Face matching and retrieval using soft biometrics, IEEE Trans. Inf. Forensics Secur. 5 (3) (2010) 406–415.
[26] R. Gonzalez, R.E. Woods, Digital Image Processing, third ed., Pearson Prentice Hall, New Jersey, 2008.
[27] P. Rosin, Measuring shape: ellipticity, rectangularity, and triangularity, Mach. Vis. Appl. 14 (2003) 172–184.
[28] T. Wallace, O. Mitchell, Analysis of three-dimensional movement using Fourier descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 2 (6) (1980) 583–588.
[29] I. Bartolini, P. Ciaccia, M. Patella, WARP: Accurate retrieval of shape using phase of Fourier descriptors and time warping distance, IEEE Trans. Pattern Anal. Mach. Intell. 27 (1) (2005) 142–147.
[30] C. Dalitz, C. Brandt, S. Goebbels, D. Kolanus, Fourier descriptors for broken shapes, EURASIP J. Adv. Signal Process. 161 (2013) 1–11.
[31] A. Folkers, H. Samet, Content-based image retrieval using Fourier descriptors on a logo database, in: Proceedings of the 16th International Conference on Pattern Recognition, vol. 3, 2002, pp. 521–524.
[32] W. Rasband, ImageJ, U.S. National Institutes of Health, Bethesda, Maryland, USA, 1997–2016, http://imagej.nih.gov/ij/.

From: [redacted]
To:
Subject: Fwd: PCAST
Date: Friday, January 6, 2017 9:27:13 AM

Follow up email from Harold Ruslander.

---------- Forwarded message ----------
From: <[redacted]
Date: Thu, Dec 15, 2016 at 6:54 PM
Subject: PCAST
To: [redacted]

Dear Mr. Lander,

Thank you for your email. Your RFI does not specifically ask for 'black box studies' as I read it, it asks more generally for information related to the use of random accidental characteristics in the examination of footwear evidence. Your published references include information that is related to the use of damage and wear features to reduce the population of possible sources of an impression that are not mentioned in the report. Perhaps, as my response suggested, more discussion with members of the practitioner community in this area of expertise could provide some insight about the relevance of these studies to your topic in question. There is at this time a black box study in progress at West Virginia University, it is anticipated that those results will be published in the next year.

The majority of work in the forensic examination of footwear involves the association of class/manufactured characteristics of footwear outsoles to impressions. The PCAST report does not question this aspect of the analysis, and therefore does not question the vast majority of the work being conducted. All features are evaluated in the same manner in the examination, it is their value that varies and that is the subject of many of the papers that were provided to you. If it is the comparison process you are questioning, it is unclear why one isolated aspect of a process is being questioned. These, and other questions may be resolved with some continued dialogue.

I believe a meeting with actual practitioners and you or your designee would be much more fruitful than speaking with me. I am not a subject matter expert in that field but would be happy to suggest one or more to you if you want.

I would also suggest that funding for face to face meetings between actual practitioners and those in your group be requested so that these issues can be fully explored, perhaps you are able to arrange such funding.

Harold Ruslander, President, The IAI