U.S. Department of Justice
Office of Justice Programs
National Institute of Justice

National Institute of Justice

Forensic Science Research and Evaluation Workshop:

A Discussion on the Fundamentals of Research Design and an Evaluation of Available Literature

May 26–27, 2015
Washington, D.C.

NCJ 250088

U.S. Department of Justice
Office of Justice Programs
810 Seventh St. N.W.
Washington, DC 20531

Loretta E. Lynch
Attorney General

Karol V. Mason
Assistant Attorney General

Nancy Rodriguez, Ph.D.
Director, National Institute of Justice

This and other publications and products of the National Institute of Justice can be found at:

National Institute of Justice
Strengthen Science • Advance Justice
http://www.nij.gov

Office of Justice Programs
Innovation • Partnerships • Safer Neighborhoods
http://www.ojp.usdoj.gov

The National Institute of Justice is the research, development and evaluation agency of the U.S. Department of Justice. NIJ's mission is to advance scientific research, development and evaluation to enhance the administration of justice and public safety.

The National Institute of Justice is a component of the Office of Justice Programs, which also includes the Bureau of Justice Assistance; the Bureau of Justice Statistics; the Office for Victims of Crime; the Office of Juvenile Justice and Delinquency Prevention; and the Office of Sex Offender Sentencing, Monitoring, Apprehending, Registering, and Tracking.

Forensic Science Research and Evaluation Workshop

A Discussion on the Fundamentals of Research Design and an Evaluation of Available Literature

Edward G. Bartick and McKenzie A. Floyd, Eds.

Forensic Science Research and Evaluation Workshop: A Discussion on the Fundamentals of Research Design and an Evaluation of Available Literature

This publication is based on a workshop funded by a National Science Foundation grant (SMA-1533843) from the Science of Science and Innovation Policy and Biological Anthropology Programs in the Directorate for Social, Behavioral and Economic Sciences (SBE), and the National Institute of Justice (NIJ), Office of Justice Programs, U.S. Department of Justice (DOJ).

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Department of Justice.

The workshop was conducted at the American Association for the Advancement of Science (AAAS), Washington, DC, May 26-27, 2015.

Edward G. Bartick, Ph.D.
Principle Investigator/Manager
&
McKenzie Floyd, B.A., M.S.
Rapporteur

The George Washington University
Department of Forensic Sciences
2100 Foxhall Road
Washington, DC 20007

May 6, 2016

Table of Contents

Overview

Edward G. Bartick

Introduction

This publication is based on a "Forensic Science Research Evaluation Workshop" sponsored by the NSF and the NIJ and was held at the AAAS headquarters in Washington, DC. The impetus for the workshop was recent criticisms of the forensic sciences from public, legal, and scientific sources. One of the more important critical reports was the 2009 National Research Council report, *Strengthening Forensic Science in the United States: A Path Forward.*[1] It was highly critical of the scientific foundations for several of the forensic disciplines, declaring that "Little rigorous systematic research has been done to validate the basic premises and techniques in a number of forensic science disciplines." and "…a statistical framework that allows quantitation … is greatly needed."[1] p.189.

Statistical frameworks for reporting the significance of physical evidence have only recently been pursued broadly. Since the early1990s, DNA analysis testimony began stating the probability of matches between question and known evidence and is the only forensic discipline where it has been routinely applied. With the statistics of DNA analysis being well addressed, researchers have begun to look at a mathematical and statistical basis for pattern comparison analysis such as latent fingerprints,[2] fired bullets,[3] and toolmarks.[4] Materials such as paints, fibers, and tapes that are often found as physical evidence at crime scenes could be of greater value should we establish a statistical significance of an association with a suspect or victim. Knowing the abundance and variation in composition of these materials could be used to establish probability estimates of these materials randomly being found at the scene. Historically there has been resistance to this approach, because information on manufactured materials has been considered to be too difficult to maintain due to production changes. However, the establishment of well-maintained centralized databases are possible and should be developed.[5]

A second critical review of the forensic sciences by Mnookin, et al.[6] reported that the forensic sciences

> …must develop - a well-established scientific foundation. This can only be accomplished through the development of a research culture that permeates the entire field of forensic science. A research culture…must be grounded in the values of empiricism, transparency, and a commitment to an ongoing critical perspective.

More recently, in January 2015 the National Commission of Forensic Sciences (NCFS) Subcommittee on Scientific Inquiry and Research[7] expressed a dire need for more rigorous standards of scientific research and resulting publications in the area of forensic science.[6]

With the foundations of forensic science under scrutiny, there is a general call to review the forensic science literature for strengths and weaknesses. This must be done by practitioners,

and forensic scientist researchers in collaboration with academic scientists in traditional disciplines. Research articles determined to be sound science should be used to incorporate analytical methodology into standards for laboratory analysis. Currently, such standards are being addressed by the National Institute of Standards and Technology's (NIST) Organization of Scientific Area Committees (OSACs)[7] and hopefully will be evidence-based.

However, the ability of today's forensic scientists to properly evaluate the forensic literature has been called into question and a need to establish a research culture to "place themselves on an appropriately secure foundation in the twenty-first century"[6] has been postulated. This criticism relates to the training forensic scientists undergo. Typically, forensic science educational programs do not offer full courses or other opportunities for students to learn research methods in any depth. Statistics is not a requirement of the Forensic Science Education Programs Accreditation Commission (FEPAC) for a Masters of Forensic Science graduate program. FEPAC requires a research project culminating in a report "suitable for publication". However, without the teaching of research methods, rigorous research is not possible. The Master's level is typically the terminal degree for practitioners in the forensic sciences. Research training, extensive investigations, and publishing are required for PhD degrees and are key to the development of a research culture. In the USA, only one recently established doctorate program currently offers a PhD in forensic science, a few universities offer interdisciplinary programs with a forensic science track, and a few programs confer PhD in forensic chemistry, forensic molecular biology, or forensic toxicology. Consequently, most practicing forensic scientists are neither equipped to evaluate the research papers of others, nor to conduct the quality of research that would be expected of an academic discipline.

Our workshop was formed to discuss the fundamentals of research design and the evaluation of the literature. The NCFS recognized that forensic scientists and the OSACs would benefit from tools to help them assess the scientific literature in order to conform to higher scientific standards in critical thinking and laboratory performance. We hope that this publication will provide some grist for evaluating and elevating the research efforts in the forensic sciences and that it may be useful to OSAC members, advanced practitioners, and peer reviewers of the basics of research and evaluation of forensic science literature, as well as for Directors of Master of Forensic Science educational programs.

Organization of the workshop

All participants were selected by a planning committee as listed in the acknowledgements. The committee decided on three subject areas, each consisting of a half-day session as follows: 1) experimental design and statistics; 2) interpretation and assessment, and; 3) policy implications. The sessions are listed as sections in this publication.

The goal was to bring together a range of 17 experts in the experimental and behavioral sciences, law, policy and government funding to address the need for a higher standard of forensic science research. Each session consisted of one plenary speaker and four to five additional speakers. Each speaker had one-half hour to present their topic. A panel discussion was held the end of each session with questions from the other workshop participants and guests.

Each participant has submitted a short essay of the topic they presented at the workshop, and those write-ups are included here in this publication. Additional observations and conclusions were made during the panel discussions and these are outlined after the write-ups of each section. The essays and discussions are listed are listed as chapters in this publication.

As a summary, there is an outline of topics to evaluate the forensic science literature. The outline provides important considerations when reviewing submitted papers for publication, planning a research project, or simply determining the scientific quality of the forensic literature. This report is intended to be a guide to plan forensic science research and assess its literature. The topics are not all-inclusive and are meant as a starting point for assessment. Each topical write-up has significant references to assist in a greater depth of background on the subject. For each particular discipline within the forensic sciences, evaluators will need a thorough knowledge base of the specific discipline to properly evaluate writings. If the evaluator is not strong in statistics, it is recommended that they confer with a statistician. A close look will be required to determine if the statistics used are appropriate.

Edward G. (Ed) Bartick is a research professor at The George Washington University Department of Forensic Sciences, Washington, DC, who is involved in the development of forensic analytical methods of evidential materials. Dr. Bartick completed a Ph.D. at the Institute of Materials Science at U. Connecticut in 1978. He has worked for pharmaceutical, instrument, and a materials production companies doing analytical development. In 1986, he joined the FBI Laboratory as a research scientist in forensic methods development. In 1991 he started a one-week class entitled “Infrared Spectrometry for Trace Analysis” for forensic examiners. Dr. Bartick has acted as research advisor for Ph.D. and M.S. graduate students from U. Virginia, Virginia Tech, and George Washington U. on forensic vibrational spectroscopy thesis projects at the FBI Academy. In January 2007, He retired from the FBI to direct the Forensic Science Program at Suffolk U. in Boston where he expanded the curriculum. He returned to the Washington, D.C., area to join GWU 2013. Dr. Bartick has authored 60 technical publications, including 11 book chapters. He was awarded the FBI Director's Award in 1994 and 1996. In 1994 he founded the Scientific Working Group for Materials Examination (SWGMAT). He chaired the group through 1997 and continued to play an active role as chair of the Database Subgroup until spring of 2014 when the Organization *of Scientific Area Committees* (*OSACs*) at NIST assumed the role of SWGs. Dr. Bartick is a Fellow of the American Academy of Forensic Sciences, a charter member of American Society of Trace Evidence Examiners (ASTEE), a member of the American Chemical Society and the Society for Applied Spectroscopy.
ebartick@gwu.edu

References

1. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press, 2009.
2. Stephen J. Taylor, Emma K. Dutton, Patrick R. Aldrich, Bryan E. Dutton Application of Spatial Statistics to Latent Print Identifications: Towards Improved Forensic Science Methodologies NCJRS Report, 2012. https://www.ncjrs.gov/pdffiles1/nij/grants/240590.pdf
3. Bacharach B A. Statistical Validation of the Individuality of Guns Using 3D Images of Bullets. NCJRS Report, 2006. https://www.ncjrs.gov/pdffiles1/nij/grants/213674.pdf
4. Petraco NDK, Chan H, De Forest PR, Diaczuk P, Gambino C, Hamby J, Kammerman FL, Kammrath BW,.Kubic TA, Kuo L, McLaughlin P, PetilloG, Petraco, NPhelps, M.S.; Peter A. Pizzola EW, Purcell DK, Shenkin P, Application of Machine Learning to Toolmarks: Statistically Based Methods for Impression Pattern Comparisons, NCJRS Report, 2012. https://www.ncjrs.gov/pdffiles1/nij/grants/239048.pdf
5. Bartick EG, Roberts K, Morgan SL, Goodpaster JV. A Statistical Approach to the Discrimination and Match Capability to Provide Scientific Basis for Estimating Significance of Fiber Association in Forensic Practice. In: Proceedings of the American Academy of Forensic Sciences Annual Meeting; 2013 Feb 18-23; Washington (DC); 90.
6. Mnookin JL, Cole SA, Dror IE, Fisher, BA, Houck, MM, Inman, K, Kaye, DH, Koehler, JJ, Langenburg, G, Risinger DM, Rudin, N, Siegel, Stoney, DA. The Need for a Research Culture in the Forensic Sciences, 58 UCLA Law Review, 725-779 (2011).
7. National Commission on Forensic Science, Scientific Inquiry and Research Subcommittee. Views Document: Scientific Literature in Support of Forensic Science and Practice. http://www.justice.gov/ncfs/file/786591/download
8. National Institute of Standards and Technology, Organization of Scientific Area Committee. http://www.nist.gov/forensics/osac.cfm

Section 1: Experimental Design and Statistics

Plenary I.
The Status of Research in the Forensic Sciences

Constantine Gatsonis

1. Scientific challenges in the multidisciplinary world of the forensic sciences

The NAS report on *Strengthening Forensic Science*[1] examined both the science and the practice in the forensic disciplines across the country. The report discussed a broad range of challenges facing the forensic science community, from disparities in resources, facilities and training across the country's jurisdictions to lack of mandatory standardization, certification, and accreditation and to political realities and evidence admissibility issues. The report also documented uneven development of the broad range of forensic disciplines and called for major emphasis on developing scientific research and educational programs in the forensic sciences.

Any examination of the forensic sciences should start by recognizing their *multidisciplinary* nature. Indeed, the more advanced forensic disciplines draw methods and expertise from a variety of scientific disciplines. For example, nuclear DNA and mitochondrial DNA analysis originated in molecular biology; and substance identification uses methods from analytical chemistry. Such forensic disciplines are generally based on solid scientific grounds because the validity of those methods has been established through past and ongoing research and development. If the analyses are executed according to the principles of science, they can be very reliable.

As an example, when a sample is matched to an individual using DNA analysis, the analysis can also provide an estimate of the probability that the sample could have belonged to another individual. This is known as the "random match probability" and is typically very small. There are many reasons why the science of DNA analysis rests on a solid foundation including: i) the extensive, peer-reviewed research behind the biological explanations for individual-specific findings; ii) the probabilities of false positives having been explored and quantified in some settings (even if only approximately); iii) the laboratory procedures being well specified and subject to validation and proficiency testing; and iv) the clear and repeatable standards for analysis, interpretation, and reporting. In contrast to DNA analysis, when a fingerprint is declared a "match" it is not yet feasible to estimate the probability that the print could belong to someone else (i.e., random match probability.) Just as concerning, examiners typically express their findings in a yes/no fashion, without reference to the error probabilities. Finally, the reproducibility of the results is different between these two types of analysis.

Beyond DNA and chemical analyses, a good number of forensic disciplines work on the identification of patterns. The analysis in these disciplines examines whether it is possible to link a pattern from a crime scene—which may be a latent fingerprint impression, markings on a spent bullet, patterns from a fire, blood-spatter patterns, and so on—with analogous patterns from a weapon, tool, finger, etc., associated with a suspect. The vast majority of these methods have been developed by the forensic science community, with little input from the broader world of science.

The NAS report notes that "The level of scientific development and evaluation varies substantially among the forensic science disciplines…[w]ide variability exists across forensic science disciplines with regard to techniques, methodologies, reliability, error rates, reporting, underlying research, general acceptability, and the educational background of its practitioners". The report then calls for research to address issues of accuracy, reliability, and validity in the forensic science disciplines. In particular, the research needs include: (a) The conduct of studies establishing the scientific basis for demonstrating the validity of forensic methods; (b) The development and establishment of quantifiable measures of reliability and accuracy of forensic analyses. The corresponding studies should reflect as closely as possible the *actual* practice using realistic case scenarios and should develop estimates of performance measures which are averaged across a representative sample of forensic scientists and laboratories; (c) The development of quantifiable measures of uncertainty in the conclusions of forensic analyses; (d) The development of automated techniques capable of enhancing forensic technologies; and (e) The conduct of studies of human observer bias and the sources of human error and contextual bias in forensic examinations. Importantly, the NAS report stresses that research in the forensic sciences should be peer reviewed and published in respected scientific journals.

2. Elements of an evaluation of the accuracy of forensic analyses.

Understanding and quantifying statistical uncertainty and the magnitude of potential error in scientific results are fundamental objectives in the sciences, including the forensic sciences. For example, laboratory analyses are subject to measurement error (i.e. uncertainty about true quantity); fingerprint analyses can lead to false identification of individual prints because of observer error or low specimen quality; DNA analyses can lead to false identification of individuals because of contamination of samples or laboratory errors.

Fuzzy use of language has been pervasive in discussions of errors and error rates in the forensic disciplines. In this section we describe basic concepts of the assessment of error in making the two common types of binary determinations in forensic analyses, *individualization* and *classification*. The former addresses the question of whether a piece of evidence can be attributed to a specific source. For example, was a particular fingerprint obtained from a specific individual? The latter addresses the question of whether a piece of evidence can be attributed to a class of sources. For example, was a piece of car paint obtained from a specific car model? At present, few forensic modalities have potential for addressing individualization questions but several of them have potential for addressing classification questions.

2.1 Accuracy for classification tasks

Studies of the accuracy of a forensic analysis to perform classification tasks can be developed using the established methods for assessing the accuracy of diagnostic tests. In such

studies units are classified by the test (in our case the forensic analysis) and by a reference standard. Importantly, the test result needs to be generated without knowledge of the reference standard. When a total of *n* units are classified by both test and reference standard, the results can be presented in Table 1. For example, a question of interest in hair analysis may be whether a particular specimen belongs to individuals from a particular group G. Thus the analysis would declare whether the specimen comes from an individual from group G or not and the reference standard would provide the definitive information on the particular individual.

Table 1.

	Forensic analysis result	
Reference standard (Truth)	"yes"	"no"
"yes" (Target condition present)	TP (True Positives)	FN (False Negatives)
"no" (Target condition absent)	FP (False Positives)	TN (True Negatives)

The accuracy of the forensic analysis can be assessed from the perspective of *detection* or *prediction*. The two perspectives are distinct and complementary to each other.

For detection, we derive two key measures of performance:

- *Sensitivity*, defined as the probability that the analysis will detect the target condition when the target condition is present.
- *Specificity*, defined as the probability that analysis will declare the target condition is not there when the target condition is absent.

In the notation of the table, the sensitivity would be estimated by the fraction TP/(TP+FN) and the specificity would be estimated by the fraction TN/(FP+TN). The corresponding measures of error are given by (1-sensitivity), estimated by FN/(TP+FN), and (1- specificity), estimated by FP/(FP+TN). The total sum of the counts in the four cells of the table equals *n.*
For prediction, we also derive two key measures of performance:

- *Positive predictive value (PPV*), defined as the probability that the target condition is present, given that the analysis result indicated its presence.
- *Negative predictive value (NPV*), defined as the probability that the target condition is absent, given that the analysis result indicated it is absent.

In the notation of the table, the PPV would be estimated by the fraction TP/(TP+FP) and the specificity would be estimated by the fraction TN/(FN+TN). The corresponding measures of error are given by (1-PPV), estimated by FP/(TP+FP), and 1- NPV, estimated by FN/(FN+TN).

An extensive literature on the design and analysis of studies of diagnostic and predictive accuracy of tests is available and can be used for the assessment of forensic modalities as well[2]. In particular the literature includes methods for quantifying the statistical uncertainty of estimates of diagnostic and predictive accuracy, assessing the potential impact of covariates on accuracy, and quantifying the extent of variation in accuracy between individual analysts and laboratories.

2.2 Accuracy for individualization tasks

The approach for assessing classification accuracy can also be used to assess the accuracy of analyses aiming at individualization. For example, in such an experiment one or more fingerprint analysts examine pairs of prints. Some of the pairs are prints from the same individual and others from different individuals. Thus each pair will be classified as "match" or "not match" by the analyst and the reference standard. The 2x2 table from an experiment involving a single rating per pair would be as in Table2.

Table 2.

	Fingerprint analysis result	
Reference standard ("Truth")	match	not match
Pair of prints comes from same individual (true match)	TP (True Positives)	FN (False Negatives)
Pair of prints comes from different individuals (true not match)	FP (False Positives)	TN (True Negatives)

Data from this type of experiment can be used to estimate error rates for fingerprint analysis using the statistical methods discussed in the previous section for classification tasks. A recent example of such a study was the evaluation of the accuracy of fingerprint analysis conducted by the FBI Laboratory.[3] Studies can also be designed to evaluate the performance of individual analysts and/or groups of analysts and to evaluate the impact of such factors as latent print quality and analyst training.

3. Experience from other disciplines

Research on the accuracy and reliability of forensic modalities can benefit from paradigms developed in other branches of science, notably diagnostic medicine and clinical chemistry. An extensive, rigorous, and on-going research enterprise underlies the practice of diagnostic medicine. This research, for example, has generated estimates of how accurate is digital mammography in identifying breast cancer, CT Colonography in identifying a suspicious polyp, and MRI in determining how extensive is a prostate cancer. Research has also assessed the influence of various factors, such as context, training, and interpretation conditions on diagnostic accuracy and also how accuracy may vary across radiologists and imaging centers.

Many of the findings from diagnostic medicine could be relevant to forensic science. In particular, variability in performance among test interpreters is commonplace in diagnostic medicine and can be strikingly large. For example, in a landmark study of the diagnostic accuracy of mammographers the sensitivity of individual mammographers interpreting *the same* set of mammograms had a range of more than 40%.[4] The "moving target" problem, created by the rapid evolution of technology may also become an important issue in forensics However, the potential to conduct prospective studies of the diagnostic performance of medical tests as practiced in everyday use may not be easily applicable to forensic analysis. In diagnostic medicine studies, subjects are enrolled prospectively and both test and reference standard information is obtained as the subjects move through the process of clinical care.[5] The design and implementation of such a prospective evaluation of accuracy using real cases would be rather challenging in the forensic disciplines.

4. Concluding remark

The NAS report on *Strengthening Forensic Science* laid out an ambitious agenda for the science and the practice of the forensic disciplines. In the aftermath of the report interest and research activity in the forensic sciences has grown, legislation has been proposed, and government initiatives have been unveiled. These are all encouraging developments, especially because the problems identified in the NAS report and the potential solutions still lie ahead of us.

Constantine Gatsonis, PhD is the Henry Ledyard Goddard University Professor and Chair of the Department of Biostatistics at Brown University School of Public Health, Providence, RI.

Dr. Gatsonis was educated at Princeton and Cornell and was elected fellow of the American Statistical Association. He co-chaired the NAS Committee on Identifying the Needs of the Forensic Sciences Community, which issued its report in 2009. He currently chairs the NAS Committee on Applied and Theoretical Statistics and is a member of the Committee on National Statistics. Dr. Gatsonis is a leading authority on the evaluation of diagnostic and screening tests and has made major contributions to the development of methods for medical technology assessment and health services and outcomes research. He is a world leader in methods for applying and synthesizing evidence on diagnostic tests in medicine and is currently developing methods for Comparative Effectiveness Research in diagnosis and prediction. Dr. Gatsonis is Network Statistician of the American College of Radiology Imaging Network (ACRIN) since the formation of the Network in 1999 and also serves as a Group Statistician for the ECOG-ACRIN collaborative group. He is lead statistician for several current and past trials, including the Digital Mammography Imaging Screening Trial (DMIST) and the National Lung Screening Trial (NLST). gatsonis@stat.brown.edu

References

1. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press, 2009.
2. Zhou X-H, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine, 2nd Ed. Hoboken, NJ: Wiley, 2011.

3. Ulery B, Hicklin R, Buscaglia J, Roberts M. Accuracy and reliability of forensic latent fingerprint decisions. PNAS 2011;108(19):7733-8.
4. Beam C, Layde P, Sullivan D. Variability in the Interpretation of Screening Mammograms by US Radiologists. Arch Int Med, 1996;156:209-13.
5. Gatsonis CA, McNeil BJ. Collaborative diagnostic evaluation studies: the RDOG experience. *Radiology*, 1990;175:571-5.

Experimental Design in the Physical Sciences

Stephen L. Morgan

Scientists are often asked to measure chemical or physical properties of materials or systems; however, proposed measurements often do not provide the required information. We find ourselves reformulating questions to be answered, and working to identify information required to answer those questions, making new measurements, and interpreting resulting data. Experimental design is at the center of basic and applied research. The objective of this paper is to provide an historical basis for the evolution of experimental design concepts and to discuss fundamental requirements for validation of designed experiments.

The scientific revolutions of the 17th and 18th century were based on planned experiments, replicated measurements, and the statistical analysis of data using mathematical modelling techniques, including least squares. By the early 20th century, systematic experimentation in agriculture to improve crop yields was commonplace. Among many other statisticians, William S. Gosset and Sir Ronald A. Fisher, were instrumental in expanding the use of statistics in scientific investigations.[1-3] Gosset, armed with Oxford degrees in both mathematics and chemistry, was hired by the Arthur Guinness Son and Company in Dublin in 1899 to introduce scientific methods, including statistics, into the practice of brewing beer. Experiments involving comparisons of malting quality depended on a number of experimental variables (factors). Because these experiments were time-consuming and the number of experiments conducted were small, Gosset realized that the uncertainty of the variability (expressed by the standard deviation) was relatively large and not well described by normal distribution theory. By 1908, he had derived the sampling distribution of the mean for the case of small sample sizes and published (with the permission of Guinness) under the pseudonym "Student."[2] Student's t-test is arguably the most widely used test of statistical inference, and was featured in R. A. Fisher's text on statistical methods for researchers.[3]

About the same time that Gosset and Fisher were explaining statistics to researchers, statistical process control (SPC) for achieving quality in manufactured products was introduced by Walter A. Shewhart.[4] He assumed that a random process is modeled by a Gaussian distribution of variation about a mean and described the control chart as a tool for discriminating between chance variation in a process, and "assignable causes" due to systematic effects. The control chart becomes more than just a scatter plot of data when the means of successive sub-groups of data (the x-bar chart) are plotted along with the resulting range of data (the range chart) in each subgroup. After about 20-30 subgroups have been plotted, control limits in the x-bar chart are set at plus or minus 3 standard deviations (3σ) about the mean of all the data, If variation in the process is random, this rule establishes a 99.7% confidence interval within which the sub-group means could be expected to fall; a similar 3σ confidence interval is calculated for the upper limit of the range of the data. When sequential measurements from a process remain within the 3σ limits, the process is said to be in a state of statistical control. Such 'control limits' have nothing to do with how we would like the process to behave, but rather tell us what can be

expected from the process. The value of statistical control in validating systems was emphasized by Grant Wernimont:

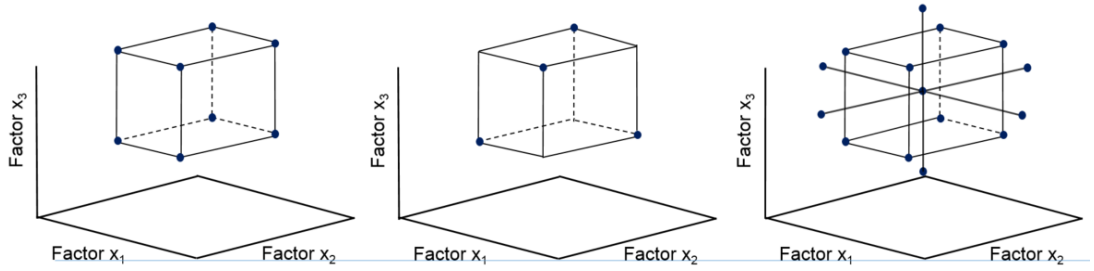
> It is unfortunate, I think, that most academic courses involving measurement do not seem to make the student aware of how important it is to achieve a state of statistical control when we set up and run a measurement process… . Measurements are not valid until we evaluate the performance characteristics of the process which produced the measurements and it is essential that the statements about the future behavior of these characteristics be correct. Statistical control is concerned with removing the assignable causes of variation in a measurement process (or correcting for their effects) so that we can associate approximate levels of confidence with these statements.[5]

Despite the value of statistical process control, improvement of systems requires more than just passive observation. Fisher's 1935 publication of *The Design of Experiments* was meant to "satisfy the requirements of sound interpretation and intelligible experimental design, and to supply the machinery for unambiguous interpretation… If the design of an experiment is faulty, any method of interpretation which makes it out to be decisive must be faulty too."[6] Rather than depending on statistical analysis to accommodate noisy, ambiguous, or poorly designed experiments, the objective of good experiment design is to avoid those issues and provide data that is reliable and defensible. Fisher was the first to point out that 'good science' often teaches experimenters to not vary everything at once, but to vary the first experimental factor (*e.g.*, temperature) to determine its effects, then while holding the first factor constant, vary the next factor (*e.g.*, pH). However, multifactor systems often exhibit interactions, in which the effect of a factor might depend on the levels of others factor.

Factorial designs, in which all experimental factors of interest are varied at all combinations of factor levels, enable detection of interactions. Just like first-order effects (slopes), interactions can be positive (synergistic) or negative (antagonistic). Computations to estimate factor effects are commonly performed now by fitting data using linear regression models. In such models, the total number of experiments must be equal to or greater than the number of parameters (p) to be estimated. Results from the 3^2 factorial design shown in Figure 1(a), can be fitted to an eight-parameter linear model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3$$

The β_0 term represents an intercept parameter; β_1, β_2, and β_3 are first-order parameters for the three factors; β_{12}, β_{13}, and β_{23} are the three two-factor interactions, and β_{123} is the three-factor interaction parameter. The one-half fractional factorial design in Figure 1(b) employs four experiments to estimate the intercept parameter plus the three first-order factor effects. Both of these designs are 'saturated' designs, in that all the degrees of freedom are used in estimating parameters ($n = p$), and the models fit the data perfectly ($R^2 = 1$). In practice, additional replicate measurements or experiments at new locations (*e.g.*, a center point) might be added.



(a) (b) (c)

Figure 1. Three common experimental designs: (a) two-level three-factor (3^2) factorial design; (b) three-factor half-fractional factorial design ($½\times3^2$); and, (c) three-factor central composite (a combination of a 3^2 factorial and a three-factor 'star" design).

Fractional factorial designs are often used for screening, with a minimum number of experiments, a large number of factors to determine which have the greatest influence on a response. Fractional factorials (without extra experiments) are used to estimate first-order parameters (or main) effects) of factors. The downside is that all screening designs have an associated confounding scheme which dictates which factors are confounded with one another. The presence of a substantial interaction effect in a study may bias estimates of other effects. Fractional factorials are often conducted with relatively wide factor levels, which 'turns up the volume' on main effects. The possibility of confounded interactions being might be ignored initially, and additional experiments done later to investigate interactions of interest. The central composite design in Figure 3(c) supports fitting a full second-order model with curvature effects in addition to all interactions. This design is widely used for exploring response surfaces and discovering optimum conditions for experimental factors.[7]

A first requirement to insure validity of experimental designs is avoidance of happenstance correlations between effects of controlled factors and uncontrolled environmental variables. For example, if the required number of experiments is too large to be performed on the same day, uncontrolled variations in the environment or in the materials under examination could introduce bias into outcomes. Suppose that all experiments at a low level of an experimental factor (*e.g.*, experiments at a pH of 3.7) are run on the first day, and all experiments at the high level of that factor (*e.g.*, experiments at pH 7.7) are run on the second day. If a day-to-day trend is present due to instrument drift, sample degradation, or some other uncontrolled change in the experimental environment, the effects of the pH factor will be inextricably mixed, or confounded, with the systematic day-to-day trend. The estimation of the pH effect is biased and invalidated by the presence of the time trend. This is where *randomization* comes into play. By randomly assigning experiments to be run on different days, the experimenter can insure that the time trend is not correlated with any factor effect. Alternatively, 'insurance can bought' against extraneous variation due to uncontrolled systematic effects by *blocking*. This involves arranging experiments in blocks that have homogenous conditions with respect to the uncontrolled variables. The experiments are usually partitioned into blocks by fractional replication which also provides an estimate of the "block effect". The potentially confounded block effect (the drift of the system between days) is the difference in mean response of the two blocks, and is not confounded with the factors effects of interest estimated from the combined data of the two blocks.[3, 7-8]

A second requirement for design validation is *replication*: the independent repetition of experiments conducted under the same experimental conditions (*i.e.*, at the same factor levels). If *n* represents the total number of experiments in an experimental design, and *f* represents the number of experiments carried out at unique locations, then $n - f$ represents the number of 'extra' replicate experiments. Replication not only allows the calculation of responses with higher precision, but also enables an estimate to be made of variability due to purely experimental uncertainty, on which a third requirement for design validation is based.

Designed experiments should provide checks for lack of fit of the model. Consider the impossibility of judging the adequacy of a straight line fitted to only two data points. Unless the number of experiments carried out at unique locations (*f*) is greater than the number of parameters (*p*), the variability due to lack of fit cannot be estimated. If $f > p$, then an F-test for lack of fit can be conducted based on the ratio of the variance due to lack fit and the variance due to purely experimental uncertainly. This test based on analysis of variance of linear models is well documented in the literature.[7-8]

A fourth requirement for assuring model validity is to check if model predictions are accurate. Inspection of residuals (differences between measured and predicted responses) might begin by looking for trends in plots of residuals (for ordinary residuals, or after scaling) against the time order of experiments, or against the levels of each factor in the design. If no replicate measurements are present, residual plots might be the only way to look for model discrepancies.[7] Most researchers are familiar with the correlation coefficient (R), or it squared value, the coefficient of determination (R^2), and interpret these statistics as "how well the model fits the data." However, R and R^2 are not goodness of fit measures. Figure 2 shows two data sets for which it is obvious, based on the discrepancies between data and models, that the line fits both sets of data equally well—the residuals are identical. The R^2 of 0.9401 means that the variation accounted for by the slope of the first model is 94.01 % of the variation in the original data about its mean; the R^2 of 0.0095 means that the variation accounted for by the slope of the second model is 0.95% of the variation in the original data about its mean.[8]

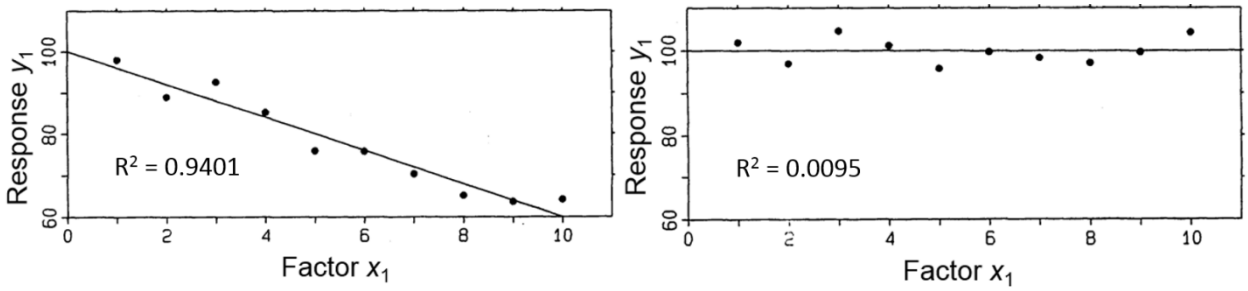


Figure 2. R^2 is not a goodness of fit statistic.

The planning of experiments, acquisition of data, and data analysis blend together in practice. Data analysis and interpretation must be done in the context of the experimental design employed, and within the limitations of both the measurements made and the data treatment. As pointed out by George Box, using designed experiments for discovery and improvement of processes and systems is not a one-shot deal; scientific investigations typically employ an iterative process in which data gathering and data analysis alternate.[9] Experimentation and data interpretation in the context of an appropriate model requires subject matter expertise aligned

with a sequential and iterative search for understanding. Data analysis also involves another iterative process in which inference from a tentatively entertained model is interpreted and subjected to critical examination by inspection of the inadequacies of the model (residuals).[9]

Finally, the impact of the "Quality revolution" sparked by William E. Deming,[10] Joseph Juran,[11] Genichi Taguichi,[12] and others in multiple countries has been dramatic. Active experimental design was recommended by Deming and Shewhart, who urged employees to improve quality with their "plan-do-check-act" cycle. Many of these concepts, including a greater awareness of statistics and experimental design, as well as an increased emphasis on quality in research have arisen from these efforts.

Stephen L. Morgan, Ph.D., is a Professor in the Department of Chemistry & Biochemistry at the University of South Carolina (Columbia, SC). He received a B.S in Chemistry from Duke University (1971), and M.S. and Ph.D. degrees (1974, 1975) from Emory University. His analytical chemistry research involves chemometrics and statistics, forensic analytical chemistry, spectroscopy, and separations. Dr. Morgan has coauthored 125 papers, 15 book chapters, and six books, including *Experimental Design: A Chemometric Approach* (1993) and *Sequential Simplex Optimization for Quality & Productivity in Research, Development, and Manufacturing* (1991). Dr. Morgan has mentored the research of 42 graduate students and 135 undergraduates. Dr. Morgan has taught, with Dr. Stanley N. Deming, over 680 short courses on experimental design and statistics for the American Chemical Society and other organizations. Awards include: Distinguished Undergraduate Research Mentor, University of South Carolina, May 2007; Sigma Xi Distinguished Lecturer, 2008-2009; American Chemical Society, South Carolina Section, 2011 Outstanding Chemist, May 2011. He is currently on the SAC Chemistry/Instrumental Analysis Scientific Area Committee, National Institute of Standards (NIST)/Department of Justice (DOJ) Organization of Scientific Area Committees (OSAC). morgansl@mailbox.sc.edu

References

1. Pearson ES, Plackett RL, Barnard GA, Ed. 'Student': A Statistical Biography of William Sealy Gosset. Oxford: Clarendon Press, 1980.
2. Student. The probable error of a mean. Biometrika 1908 Mar; 6(1): 1-25.
3. Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd, 1925.
4. Shewhart WA. Statistical Method from the Viewpoint of Quality Control. New York: Dover Publications, Inc., 1986.
5. Wernimont G. Statistical Control of Measurement Processes. In: Devoe, JR, editor. Validation of the Measurement Process. Washington, DC: ACS Symposium Series 63, American Chemical Society, 1977; pp 1-29.
6. Box GEP, Hunter JS, Hunter WG. Statistics for Experimenters. 2nd ed. Hoboken: John Wiley & Sons, 2005.
7. Deming SN, Morgan SL. Experimental Design: A Chemometric Approach. 2nd rev. ed. Amsterdam: Elsevier Science Publishers, 1993.
8. Box GEP. Improving Almost Anything. Hoboken: John Wiley & Sons, 2006.
9. Deming WE. Out of the Crisis. Cambridge, MA: Massachusetts Institute of Technology, 1986.
10. Juran JM. Juran's Quality Handbook. 5th ed. New York: McGraw-Hill, 1999.

11. Taguchi G. Introduction to Quality Engineering: Designing Quality into Products and Processes. Tokyo: Asian Productivity Organization, 1986.

Experiments in the Social Sciences

Dietram A. Scheufele

Experiments are often thought of in connection with the physical sciences, but they also play a crucially important role in the behavioral sciences. Given their high degree of external validity (or generalizability), large-scale public opinion surveys are typically preferable when it comes to measuring attitudes among large, difficult-to-observe populations of respondents. Cross-sectional surveys have drawbacks, however, when it comes to internal validity, i.e., the ability to establish a clear causal link between an independent and dependent variable.

This essay discusses the advantages of random-assignment lab experiments in behavioral research, especially with respect to establishing causal links between variables. It focuses on (a) the basic principles and terminology of experimental research; (b) quality criteria for experimental work; (c) the new realities of large-scale survey-experiment hybrids in online settings that maximize both internal and external validity; and (d) considerations related to Institutional Review Board (IRB) protocols and other ethical issues that arose from a number of highly publicized A/B experimental studies conducted by Facebook and other social media firms in collaboration with "embedded" academic researchers.

Social scientific research, more broadly, adds insights into forensic science in at least two ways. First, a variety of social science methods – including random-assignment experiments – can provide valuable insights in legal proceedings into if and how "reasonable consumers," for instance, are able to detect forgeries, fraud, or other potentially misleading information that is being litigated. Second, social science provides systematic assessments of how and in which areas forensic science has been able to influence court proceedings, jury deliberations, and the outcomes of criminal investigations, more broadly.

Basic Principles

All evidence-based social scientific research follows a similar set of guiding principles. Three of them are particularly worth highlighting in this context. The first principle is **falsifiability**, which refers to the idea of setting up hypotheses that can be tested by rejecting (with some certainty) the possibility that observed variations are simply due to random chance. This is only possible if a second principle is followed: the idea that all research **systematically** follows a set of pre-defined procedures and methods. All respondents or participants, for example, are asked the exact same questions or are exposed to the same protocols in order to ensure that any observed variation in their behaviors or attitudes is a function of differences between respondents rather than of inconsistencies in how their attitudes or behaviors were measured. A final principle is **intersubjectivity**, i.e., the notion that researchers describe their procedures, data collections and other logistics in a way that allows other scholars to replicate any given study and see if findings hold across different samples.[1]

In the case of experiments, all of these principles allow researchers to establish internal validity – or causality – i.e., the idea that variations in the dependent variable (an outcome) are due to an independent variable (an experimental stimulus) and not due to any other influences.

There are three main criteria for establishing causality: (1) The independent variable needs to precede the dependent variable in time, i.e., the cause occurs before the effect; (2) there is some form of statistical covariation between the independent and dependent variable (measured through a correlation or other measure of association); and (3) there are no other variables that can explain away the relationship between independent and dependent variables.

Random and Control Groups

The key advantage of lab-based experimental designs over survey-only designs is the fact that they allow social scientists to establish causality across all three criteria. Cross-sectional surveys, in particular, tend to be problematic in this respect since they cannot establish time order between variables. Even though a researcher might, for example, be able to statistically link measures of the frequency with which respondents read newspapers to their knowledge about current affairs, he or she will not be able to clearly establish a causal link between the two in a cross-sectional survey design. It might be the case that more knowledgeable respondents also read newspapers more frequently. But the reverse could also be true, i.e., more frequent readers learn more about current events from the newspaper.

Experimental designs allow researchers to directly test which one of these interpretations is the correct one. Although researchers can draw on a myriad of different experimental designs,[2] most lab experiments share at least one key characteristic: Participants are randomly assigned to a treatment and a control condition. Random assignment is designed to make the control and experimental group as equivalent as possible, not just in terms of demographic characteristics, but also in terms of variables relevant to the research questions.

Both groups then undergo the exact same, systematic experimental procedure. The only difference between the experimental and control group is that the former is exposed to a stimulus designed to change behaviors, cognitions or attitudes, whereas the latter serves as a placebo condition.

Let us assume, for example, that a researcher is interested in measuring the impacts of political attack ads on voter turnout.[3] Participants in the experimental condition watch television programming interrupted by political ads with explicit attacks on a political opponent. The control group is also exposed to political ads in order to ensure that the act of watching political commercials is not what causes the effect. However, the control group sees ads without explicit attacks on the opposing candidate. As a result, the only difference between the experimental group and the control group is the *nature* of the ads that the two groups see, and any difference in attitudes or behaviors in the post-test are due to the stimulus and only to the stimulus.

What Experiments Can and Cannot Do

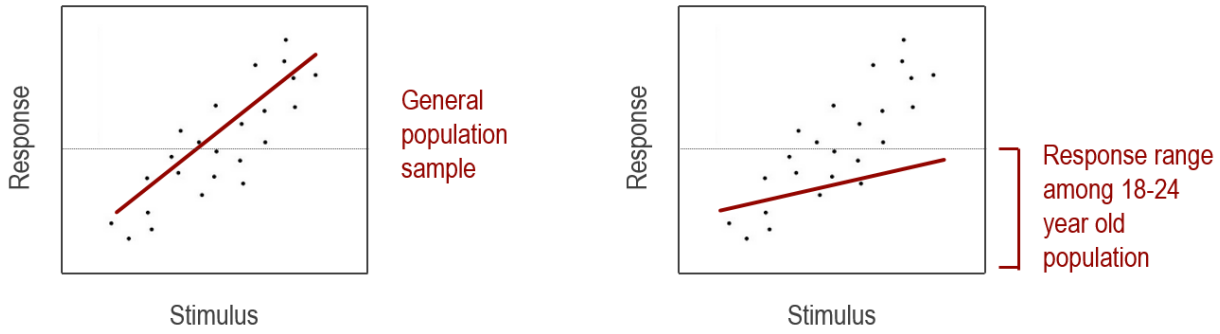
Many experiments rely on non-representative student samples, often administered through large-scale subject pools in psychology or other social science departments. In other words, they are based on convenience samples, drawn from highly-specialized populations of

college students, rather than probability samples,[4] that would allow researchers to generalize from a small subsample to the general population.

As a result, one concern often raised in this respect relates to the lack of generalizability from student samples to the broader population. In some respects, however, this criticism is misses the point. As outlined earlier, laboratory experiments in the social sciences are typically concerned with examining differences between randomized control and experimental groups, with the latter being exposed to a manipulated stimulus. As a result, they are particularly useful for establishing internal validity or causality. Laboratory experiments, in particular, are therefore less concerned with external validity or generalizability. This does not mean, however, that the reliance on student samples in experimental research does not have potential drawbacks. Two are particularly worth highlighting.

The first concern relates to **censored variables**. Samples of college students are characterized by highly constrained demographic variables, typically including a very narrow range in age and educational attainment, but also in many attitudinal or behavioral measures. In other words, college students often do not display the full range of attitudes observed in the general population but cluster at one end of a given scale. Figure 1 illustrates this problem. The plot on the left shows the "real" effect of the stimulus on an outcome variable, e.g., a person's likelihood of voting in elections, among the general U.S. population. The plot on the right shows how that relationship changes if the population from which the experimental and control group are drawn is censored, i.e., does not display the full range of attitudes or behaviors. In our example, college students might show a lower likelihood of voting overall and therefore showing a more attenuated response to the stimulus.

Figure 1: Problems arising from a censored dependent variable



As a result, it is possible to over- or underestimate effects in experiments with student-only samples, relative to what the real effect size would be if the experiments had recruited participants from the general population. In order to avoid these potential biases, a number of researchers use financial or other incentives to recruit participants from the general population and have cut back on using student-only samples.

A second concern of laboratory experiments in the social sciences relates to the idea of **captive audiences**.[5] Most experiments rely on the idea that participants are assigned randomly to the experimental and control group, and that the way that the experiment is administered is

perfectly equivalent across the two groups. As a result, researcher can be certain that any difference observed between the two groups is due to the stimulus and only to the stimulus.

One side effect of using a highly standardized procedure across respondents, however, is that participants are asked to behave in ways that might not correspond to their normal behavior in everyday life. Our earlier example of measuring the impacts of political attack ads on voter turnout is a good illustration. Many television viewers step out during commercial breaks, fast-forward through commercials on their DVRs, or even switch channels. Forcing them to watch these ads in an experimental setting, therefore, makes them a captive audience for a stimulus they may never pay attention to in the real world. As a result, lab experiments often measure the types of effects that are *possible* if participants are exposed to a message, rather than the effects that are *likely* in the real world.

The Future of Experiments

In order to work around some of the constraints of lab-based experiments, two trends have emerged in experimental research. The first trend are **population-based survey experiments**.[6] These are large-scale representative surveys that use random assignment for part of the survey questionnaire to expose respondents to particular messages,[7,8] types of question wording,[9,10] or other types of stimuli. Since all the questions before and after the stimulus are administered consistently to the complete sample, any difference between the randomly-assigned groups in their responses can be attributed uniquely to the stimulus. Most importantly, by combining a representative population survey with a random assignment experiment, population-based survey experiments allow researchers to maximize both the generalizability (or external validity) of their design and their ability to make causal claims (i.e., internal validity). Of course, the popularity of population-based survey experiments among researchers is somewhat tempered by the fact that they are significantly more expensive than simple lab-based experiments.

A second trend is what has been referred to A/B testing, e.g., randomized assignments of users to slightly modified versions of a given online platform. These A/B studies are often conducted by the companies running these platforms and legally protected by liberally-worded opt-in user agreements.

More recently, however, some **online A/B experiments** have raised significant ethical and legal issues. One example is a recent study conducted by Facebook researchers that manipulated users' news feeds without their knowledge. While this is kind of covert testing is fairly typical for corporations, this study raised concerns for at least three reasons.

First, the Facebook researchers secretly changed users' news feeds to in order to trigger positive and negative emotions, and some observers were deeply troubled by the potential risks for unsuspecting participants with a history of depression, for instance.

More importantly, however, Facebook researchers published these findings in the *Proceedings of the National Academy of Sciences (PNAS)* in collaboration with university-affiliated researchers, but without many of the safeguards that university researchers are required to put in place before collecting, analyzing or publishing data.[11] As *PNAS* Editor-in-Chief Inder Verma put it: "Obtaining informed consent and allowing participants to opt out are best practices

… under the U.S. Department of Health and Human Services Policy for the Protection of Human Research Subjects … [It is] a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out."[12]

Finally, some researchers have raised concerns about such collaborations between what they call "embedded" academic researchers and commercial research departments, given their potential for violating some of the principles of social science research introduced earlier, including intersubjectivity, i.e., the ability of other researchers to replicate (and potentially falsify) any given study. Computer scientists Derek Ruths and Jürgen Pfeffer summarized this problem best: "The rise of 'embedded researchers' (researchers who have special relationships with providers that give them elevated access to platform-specific data, algorithms, and resources) is creating a divided social media research community. Such researchers, for example, can see a platform's inner workings and make accommodations, but may not be able to reveal their corrections or the data used to generate their findings."[13] The increasing availability of Big Data and the temptation for researchers to access cheap pre-collected data from corporations, such as Facebook, will only exacerbate the ethical and legal challenges for experimental researchers in the coming years.

Dietram Scheufele, Ph.D., is the John E. Ross Professor & Vilas Distinguished Achievement Professor, at the University of Wisconsin-Madison & the Morgride Institute for Research, and a Visiting Scholar at the Annenberg Public Policy Center, University of Pennsylvania. Dr. Scheufele currently co-chairs the National Academies' Roundtable on Public Interfaces of the Life Sciences. Based mostly on representative population surveys and randomized (lab) experiments, Scheufele's research examines the role of media and other emerging modes of communication in our society. He is a fellow of the American Association for the Advancement of Science and the Wisconsin Academy of Sciences, Arts & Letters, and a member of the German National Academy of Science and Engineering. Scheufele has been a tenured faculty member at Cornell University and a Shorenstein fellow at Harvard University. His consulting experience includes work for PBS, the World Health Organization, and the World Bank.

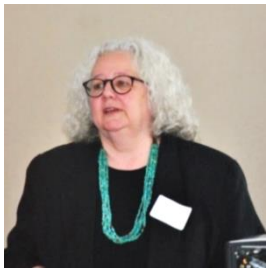
References

1. Popper K. Logik der Forschung [The Logic of Scientific Discovery]. 10th ed. Tübingen, Germany: Mohr, 1994.
2. Cook TD, Campbell DT. Quasi-experimentation: Design & analysis issues for field settings. Boston: Houghton Mifflin, 1979.
3. Ansolabehere S, Iyengar S. Going negative: How political advertisements shrink & polarize the electorate. New York: The Free Press, 1995.
4. Trochim W. The research methods knowledge base. 2nd ed. Cincinnati: Atomic Dog Publishing, 2000.
5. Hovland CL. (1959). Reconciling conflicting results derived from experimental and survey studies of attitude change. Am Psychol 1959; 14: 8-17.
6. Mutz DC. Population-based survey experiments. Princeton: Princeton University Press, 2011.

7. Anderson AA, Brossard D, Scheufele DA, Xenos MA, Ladwig P. The "Nasty Effect:" Online incivility and risk perceptions of emerging technologies. J Comput-Mediat Comm 2014; 19(3): 373–387.
8. Brossard D, Scheufele DA. Science, new media, and the public. Science 2013; 339(6115): 40-41.
9. Moy P, Scheufele DA, Eveland WP, McLeod JM. Support for the death penalty and rehabilitation: Question order or communication effect? J Appl Soc Psychol 2001; 31(11): 2230-2255.
10. Zaller J, Feldman S. A simple theory of survey response: Answering questions versus revealing preferences. Am J Polit Sci1992; 36(3): 579-616.
11. Fiske ST, Hauser RM. Protecting human research participants in the age of big data. P P Natl Acas Sci USA 2014. doi:10.1073/pnas.1414626111
12. Verma IM. Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks. P Natl Acas Sci USA 2014; 111(29): 10779. doi:10.1073/pnas.1412469111.
13. Ruths D, Pfeffer J. J. Social media for large studies of behavior. Science 2014; 346(6213): 1063-1064. doi:10.1126/science.346.6213.1063.

Non-Experimental Research in Anthropology

Marilyn R. London and Kevin G. Hatala



The scientific method is the gold standard for doing research. This method involves developing a hypothesis to answer a question, designing an experiment to test the hypothesis using experimental and control samples, collecting data from the experiment, and analyzing those data to reach a conclusion. This standard is useful and frequently necessary to develop strong theories that can be applied to understanding the world.

However, there are many conditions where some part or parts of the scientific method cannot be applied. In these cases scientists use approaches that are considered to be non-experimental research. These approaches, however, can provide valid and supportable conclusions, as discussed below.

Non-experimental research includes research where no control group is available, for a variety of reasons. The sample size available may be too small to represent a population. The researchers may have limited access to the context of the data. Often, in medicine and human biology, ethics prevents us from developing an experiment where one group of humans, for example, is given a disease or a treatment and the other is not. There may be no population from which to select controls for comparison.

Data may be collected for these types of research in a variety of ways. A survey or questionnaire may be developed, with either specific questions such as age, place of birth, or number of pregnancies; open questions, where the respondent may answer with self-generated information or opinions; or ranked questions such as those where the respondent indicates satisfaction on a scale. Observation may provide descriptive documentation, such as the number of times a child interacts with his mother in a given time frame. Researchers may be looking for patterns or basic information, such as the average age of onset of a particular disease. Case histories, both current and historic, may be studied as a group, with no controls. Archival information from sources such as medical records, or meta-analysis of data from previously published literature, can also provide valuable data for analysis, with no controls. The conclusions from the analysis of these data may be inductive rather than deductive, because the rules guiding experimental research cannot be applied.

The general public regards anthropologists from all subfields (archaeology, cultural anthropology, linguistics, paleontology, medical anthropology, primatology, forensic anthropology, etc.) as scientists who collect data by doing field research. Anthropologists analyze and interpret the data they collect to explain human variation, behavior, migration, and evolution, among other things. For the reasons listed above, anthropologists are sometimes unable to do experimental research.

For instance, archaeologists usually study people for whom there is no written history. The evidence left behind by these people may include man-made objects for everyday use, such as pottery, basketry, and hearthstones. Evidence of architecture and roads and tools of various types may be present. Human skeletal remains or mummies may be discovered. There is information in this evidence that will give indications about how the people interacted with their environment, what skills were developed, and what other populations were related to them. The human remains may provide information about health, life expectancy, and migration. However, no archaeological site preserves all the information about the people who lived there, and in many instances, even the evidence present is damaged or incomplete. There is no way to go back in time to observe these people, but comparisons may be made to populations who lived in similar climates, or who made similar artifacts, or even to modern people whose culture somehow resembles the archaeological evidence.

In forensic anthropology, a subfield of physical anthropology, each case has a sample size of 1. Although we can compare individual skeletons to others with respect to age indicators, sex, stature, ancestry, trauma, and various other identifying features, we are dealing with unique biological organisms, one at a time. Skeletal biologists do study populations of known individuals to determine what is "normal" and how much variation is present, how male skeletons differ from female, which bone measurements can provide the best estimate of stature, and indicators of other aspects of a person's biological profile, but we are often left with questions because individuals may have traits that are unique. However, working with other forensic scientists, we must do our best to identify the individual and to determine why he or she has come to us as a forensic case.

There are further limitations in forensic anthropology. The bones do not indicate how an individual identified himself or herself. A person with mixed ancestry may choose to associate primarily with individuals from one of those ancestral populations, but the bones will not indicate which group (ethnic, cultural, or racial, for example) was chosen. Individuals who are born one sex and choose to change to another cannot change their skeletons once they have reached adulthood. Information from legal documents that we use for comparison to our analyses, such as driver's licenses, include self-reported information about stature and weight which may not be accurate. The databases of measurements and observations on known individuals do not include samples from every human population.

Standards for best practices and valid techniques have been developed in forensic anthropology, as they have been in other scientific fields. However, these standards cannot anticipate every case – its context, how complete the evidence is, time and resource limitations, and the like – and are thus imperfect. Experimental research cannot be applied under these circumstances.

A Case Study from Paleoanthropology

As described above, non-experimental research is defined as research that does not directly involve controlled experiments. It is an approach that is typically applied in cases where certain variables cannot, or should not, be manipulated.

The majority of research in paleoanthropology, including our own research on the evolution of human anatomy and locomotion, involves non-experimental approaches. These are necessary because we can never go back in time and see what our fossil ancestors looked like, how they moved, and in what kinds of environments they lived. Consequently, we are never going to know anything *absolutely* by studying the fossil record. We cannot design modern experiments to *directly* test hypotheses regarding fossil organisms because modern animals and modern environments are vastly different today than they were millions of years ago. This does not mean that hypotheses cannot be tested and important questions cannot be answered, it simply means that there are some important considerations to make when designing this type of research.

To draw on a familiar reference from popular media, one way to consider the various components of non-experimental research is to deconstruct a research design into what Donald Rumsfeld would call "known knowns", "known unknowns", and "unknown unknowns". The "known knowns" are variables whose values can be known absolutely, the "known unknowns" are variables whose values are not known but can be estimated, and the "unknown unknowns" are variables whose values are not known and likely can never be known. As an example, consider the question of how human anatomy and locomotion have evolved. Looking at the data available from the fossil record, there are certain "known knowns" such as the appearance of fossils, their morphologies, and their geological context or age. The "known unknowns" would be a fossil individual's pattern of locomotion – we will never know exactly how that individual moved but we can build informed estimates based upon experimentally-determined relationships between skeletal morphology and its function. Finally, there are a number of things that we cannot know and cannot really estimate. In almost every fossil discovery, some (often several) parts of the skeleton are missing so those data will be forever unknown to us. Similarly, we can never know much about the soft tissue (muscles, ligaments, etc.) that held the skeleton together because these parts of the anatomy rarely, if ever, fossilize. One must be cognizant of each of these sets of variables, and how they influence the ability to test certain hypotheses or restrict the confidence of certain hypothesis tests.

We take all of this into consideration and adjust our approach accordingly in our recent analyses of 1.5 million year old fossil human footprints that we have discovered and excavated in northern Kenya.[1] In an area along the eastern shore of Lake Turkana, we have found six sites that together preserve more than 100 fossil human footprints created by more than 15 different individuals. These might not seem like huge numbers but compared with the 'typical' fossil discovery that may consist of a handful of bones or teeth from one individual, this data set is extremely large. This kind of discovery is also rare because the age of these footprint surfaces is very clear. All of these sites fall within a sandwich of three volcanic tuffs that have each been accurately dated to around 1.5 million years ago. Based on the overall size of these footprints, and what we know about the body sizes of different human ancestors and relatives living in northern Kenya at 1.5 million years ago, we believe that these footprints were most likely made by individuals belonging to the species *Homo erectus* or *Paranthropus boisei*. We know extremely little about foot anatomy and foot function in either of these species though, so these footprints offer us the first opportunity to study the foot anatomy and locomotion of our fossil relatives at 1.5 million years ago. The implications of these analyses are very important because we know that other aspects of postcranial anatomy in these species took on a dramatically more

human-like form compared with earlier human relatives such as members of the genus *Australopithecus*. These footprints offer us the first opportunity to directly assess the humanness of the locomotor patterns of 1.5 million year old hominins.

Interpreting the locomotor patterns of these fossil hominins, as preserved in their footprints, has required a non-experimental approach. We cannot find a modern taxon with an anatomy identical to that of our 1.5 million year old relatives and simply study how they move and how their locomotor pattern is reflected in their footprints. Instead, we have had to carefully construct a research design that will allow us to interpret functional patterns from these fossil footprints while acknowledging the limitations and uncertainties that are inherent to this type of fossil analysis.

To start, we can again break down our research design into constants ("known knowns"), variables that can be estimated ("known unknowns"), and variables that cannot be estimated ("unknown unknowns"). The constants include the geological age of the footprints, their number, their sizes, and their morphologies. The variables that we wish to estimate include aspects of the locomotor patterns reflected by the footprints, such as traveling speed or limb motion patterns (kinematics). And the variables that we know we can never estimate include certain details of foot anatomy (e.g., aspects of soft tissue structures) and whether or not these footprints are 'typical' of the species that lived 1.5 million years ago.

Despite the fact that our overall approach and research question are non-experimental, we have used some experimental approaches within that broader framework. Specifically, we have used experimental studies to link the data that are preserved in footprints ("known knowns") to certain aspects of locomotor patterns that we wish to estimate from fossil human footprints ("known unknowns"). For example, we have conducted experiments to determine the quantitative relationships between footprint stride lengths and traveling speed.[2] We have also examined the quantitative links between a footprint's 3-dimensional morphology (topography) and lower limb joint motion patterns.[3] In these experiments, our goal has been to uncover systematic/mechanical relationships that will allow us to use the data preserved in fossil human footprints to generate informed estimates of the locomotor patterns that produced those prints In effect, we are working to develop 'standards' by which certain biomechanical variables can be interpreted from human trackways and, in turn, fossil human trackways. One can imagine a case in which these or similar standards could be applied in a forensic setting, to determine the nature of an individual's gait from footprints at a crime scene. By applying this type of approach in a paleontological context, we have been able to develop hypotheses with robust experimental support even though they are constructed within a broader non-experimental research design.

Ultimately, our analyses have shown that the 1.5million year old footprints we have uncovered in Kenya are morphologically similar to the footprints of modern habitually barefoot people.[1] Drawing upon our experimental work, which has linked particular aspects footprint morphology to specific biomechanical variables, the morphologies of these fossil footprints suggest that the fossil humans who created them 1.5 million years ago used a walking gait remarkably similar to the one we use today. We can never know, however, whether certain details of foot anatomy in these fossil humans may have been different from our own and could have led to functional differences that are not obviously reflected in their footprints.

Furthermore, we cannot be sure that a human-like walking pattern was 'typical' of all fossil humans who lived around 1.5 million years ago. We can only state that the footprints we have analyzed suggest that a human-like walking gait was used by the particular set of individuals who produced those prints 1.5 million years ago.

As in any non-experimental research, it has been critical that we remain cognizant of what we know, what we can estimate, and what we can never know. By doing so, we are able to conduct robust hypothesis tests but, just as importantly, also generate appropriately cautious conclusions.

Marilyn London, M.A., is a lecturer in anthropology at the George Washington University and University of Maryland College Park. Ms. London is a skeletal biologist concentrating in paleopathology, the study of how diseases have affected humans throughout history and prehistory. She serves on the U.S. Department of Health and Human Services D-MORT Team where she is a forensic anthropologist dealing with the application of anthropological techniques and knowledge to incidents that include transportation accidents, hurricanes, and terrorist attacks. She has worked with the Department of Anthropology at the Smithsonian Institution, analyzing human skeletal remains to develop biological profiles and life histories on hundreds of individuals from their bones. Her fieldwork includes prehistoric and historic sites in New Mexico, Maryland, and New York City. She has served as a forensic anthropology consultant to the Office of the Medical Investigator in New Mexico and the Office of the Medical Examiner in Rhode Island. mlondon@umd.edu

Kevin G. Hatala, Ph.D., is a National Science Foundation Social, Behavioral, and Economic Sciences Interdisciplinary Postdoctoral Fellow who collaborates with a variety of paleoanthropologists, evolutionary biologists, and computer scientists from the Max Planck Institute for Evolutionary Anthropology, the American Museum of Natural History, Brown University, and the University of Colorado-Boulder. He is a paleoanthropologist interested in the evolution of human anatomy and locomotion, and their dynamic relationships with human ecology and behavior. Dr. Hatala uses a variety of experimental methods to study locomotion in modern humans and nonhuman primates in order to better understand how our anatomies influence the ways that we move. He also conducts paleontological fieldwork in northern Kenya, to uncover fossils that will advance our understanding of human evolutionary history. His postdoctoral research fellowship is aimed at developing new experimental and quantitative methods that can be used to extract anatomical and functional data from footprints in the human fossil record. His approach combines ongoing paleontological excavations of fossil human footprint sites in Kenya with laboratory experimental projects aimed at modeling the ways in which detailed aspects of anatomy and locomotion are recorded in footprints.
kevin.g.hatala@gmail.com

References

1. Hatala KG, Richmond BG. Earliest direct evidence of modern human-like foot function from 1.5 Ma hominin footprints at Ileret, Kenya. Am J Phys Anthropol 2015; 156; 160.
2. Dingwall HL, Hatala KG, Wunderlich RE, Richmond BG. Hominin stature, body mass, and walking speed estimates based on 1.5 million-year-old fossil footprints at Ileret, Kenya. J Hum Evol 2013; 64: 556-568.

3. Hatala KG, Wunderlich RE, Dingwall HL, Richmond BG. Interpreting locomotor biomechanics from the morphology of human footprints. J Hum Evol. In Press.

Appropriate Statistics

Joseph B. Kadane

Abstract

The statistics that most illuminate a given question depends first on the question, and second on the data that may shed light on it. Therefore, there is no single method appropriate to all occasions. There are, however, some distinctly different styles, of which I discuss three. These are a) data analysis, a free-form look at the data, often using various visualizations to get a feel for the data, b) classical statistics, including tests of hypotheses, confidence intervals, etc., and c) Bayesian statistics, including choice of models and prior distributions. I describe each method, elaborate on the appropriate circumstances to use them, and discuss their advantages and disadvantages.

1. The Basics

The first fundamental question is “What are you trying to find out?" Without a definite target, no analysis is likely to be satisfying. For the purpose of illustration, suppose the goal is to measure the proficiency of a particular fingerprint examiner.

The second question to address is what may already be known about the question of interest, in this case the proficiency of the fingerprint examiner. Let’s suppose that the examiner is quite new to the job. We have hopes, but this test will be our main source of information.

The third question is whether data are available to address the issue. If they are available, how did they come to be available? If they come from an experiment, how was that experiment designed? What selection effects may have limited the available data?

Selection effects are particularly important, and can be subtle. There’s the well-known story of the general in World War II who wanted to armor his fighter planes where they were being shot. But the data was from the planes that came back, and the desired inference was for the planes that did not. Therefore, it was more sensible to armor the areas of the planes that did not have bullet holes in them.

An example in forensic science involves attempts to use the presence of root banding of hairs to establish a lower bound on time since death. But the data came from hairs removed at autopsy after refrigeration. Autopsies don’t happen instantly, and refrigeration is known to slow the development of root banding, so data from autopsies is likely to exaggerate the time required for root banding to develop after death.[1]

In our hypothetical example, we’ll suppose that ten known matches of pairs of fingerprints were slipped into the workload of the examiner, and that ten known non-matches were also slipped in. In this scenario, the main selection effect is how similar the non-matched pairs are to each other, and the quality of the prints of the matched pairs. This is like a classroom

examination, in which one has to judge how hard the exam is, in order to understand how well a student did.

The next step is to look at the data, preferably by finding a way (or several ways) to plot them. Some of the standard methods of plotting data are box plots, stem-and-leaf plots, and scatter plots. It is particularly important to examine the data for outliers. Outliers might be evidence of errors in recording the data, but might also be bearers of vital information for the question of interest.

Another question is whether there are missing data, and, if so, how to think about whether the very fact that they are missing has information about the questions of interest. For example, if high-school students fail to take a national proficiency exam, might that be evidence that the student was predicted not to do well, and was discouraged from taking the exam.[2] If missing data are informative, then a full accounting of the data requires modeling the effect.

To continue our hypothetical example, lets suppose that our fingerprint examiner got 9 out of 10 true matches correct, and 8 out of 10 non-matches correct. With data this simple, there's not a lot to be done to visualize the data, but bar-plots, shown in Figure 1, at least display the data.

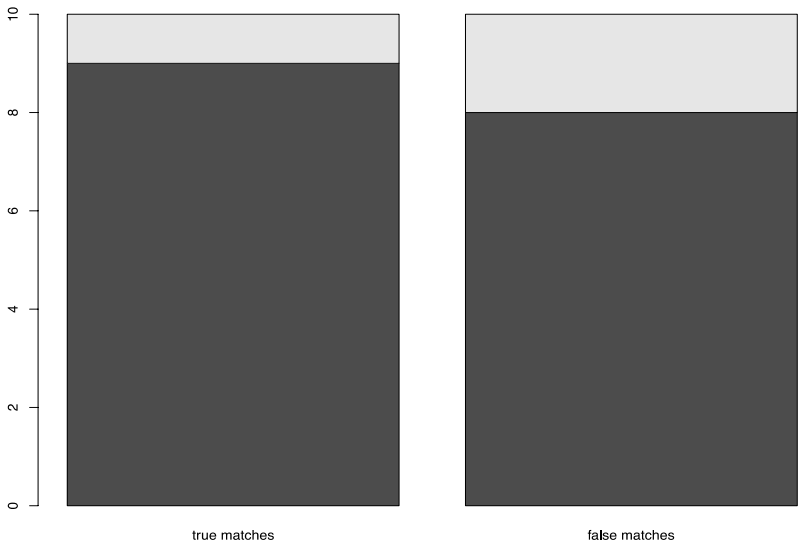


Figure 1: Bar Plot for the raw data: shaded area represents correct answers, unshaded area represents wrong answers.

2. Inference the way we're taught (Sampling theory)

The most heavily used sampling theory method is to test a null hypothesis, often at the 5% level. Suppose we think of the hypothetical data as coming from a binomial distribution, which means that it is the sum of independent "coin flips" with some probability p of coming up "correct." We might want to test the hypothesis

$$p = 1/2. \qquad (1)$$

There are several ways this might be done. A natural way in this problem is to reject the

null hypothesis if the analyst's number of errors is sufficiently small. Under the hypothesis that p=1/2, each sequence of ten results (correct or incorrect) is equally likely. There is one configuration with no errors, ten with one error, and 45 with two errors. Therefore, the probability of having one or fewer errors under the null hypothesis is

$$(1+10)/2^{10} = 11/2^{10} = 0.017 \qquad (2)$$

Since this number is less than 0.05, a sampling theory statistician would reject the null hypothesis at the 0.05 level.

Similarly, the probability under the null hypothesis, the probability of 0,1 or 2 errors is

$$(1+10+45)/2^{10} = 56/2^{10} = 0.0546$$

Since this number is greater than 0.05, a sampling theory statistician would not reject the null hypothesis in this case.

A related sampling theory technique is to report a confidence interval for p. In these two examples, the 95% confidence intervals for p are respectively (0, 0.3942), for one error, and (0, 0.05069) for two. These are the sets of null hypotheses that, had they been tested at the 0.05 level, they would not have been rejected. Observe that for 1 error, the upper limit, 0.3942 is less than 1/2, so the null hypothesis that p=1/2 would be rejected. By contrast, with two errors, the upper limit is 0.5069, greater than 1/2, so the null hypothesis p=1/2 would not be rejected.

What does all this mean? We are comparing the probability of seeing a number of errors equal to or less than those observed, to an arbitrary number 0.05. So what?

Experienced statisticians know that whether data pass or fail a test of a null hypothesis is more driven by the sample size than by how far from true the null hypothesis is. And we have easy access to the sample size without the need for calculations.

Secondly, there's the question of how to interpret such a test. According to R.A. Fisher, who advocated these methods and proposed the level 0.05, either the null hypothesis is false or something unusual has happened. But the method does not say which of these is the case, nor does it permit the calculation of a probability as to which occurred.

Furthermore, the test of significance offers no guidance about what to conclude if the null hypothesis is not rejected. It doesn't mean that the null hypothesis is true. I would interpret failure to reject as indication that the sample size was too small.

The interpretation of confidence intervals is equally fraught. It is inviting to interpret them as indicating, for example, 95% probability that the parameter (p in the example) lies in the interval. But that's not what a confidence interval is. Rather it is a statement that if I used the procedure I used to create this interval many times, in 95% of those times the interval (different every time, in principle) would contain the parameter. Thus if I have a random device that comes up heads 95% of the time, and associate "heads" with the whole real line and tails with the empty set, such an interval will cover the parameter 95% of the time. Therefore, it is a valid 95% confidence interval (and is worthless).

These difficulties with sampling theory, when one goes beyond the rote language and looks at what phrases like “significant" and “confidence" really mean, suggest that a different paradigm might better associate the calculations with the intended inferences. There is such a framework, as introduced in the next section.

3. Bayesian Statistics

The basic idea in Bayesian Statistics is to use probability as a language to quantify uncertainty. (Bayes Theorem itself is just an uncontroversial result in probability theory, having to do with conditional probability.)

Bayesians generally take a personalistic view of probability – my opinions (as expressed in probabilities) need not be the same as yours.

Usually Bayesians express their opinions in terms of a likelihood.

$$p(x \mid \theta)$$

(where $\boldsymbol{x}$ is the data and θ are the uncertain quantities of interest), and an opinion about θ,

$$p(\theta)$$

Then the model for data and uncertain quantities jointly is

$$p(x \mid \theta)p(\theta)$$

The posterior distribution after seeing data x is

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta) = \frac{p(x \mid \theta)p(\theta)}{\int p(x \mid \theta)p(\theta)\,d\theta}$$

Much of the technical work in Bayesian computing has to do with computing the integral, or working around it, as in Markov Chain Monte Carlo (MCMC).

To continue with the example, suppose that your prior distribution on p is uniform on (0,1) in both cases, and that your likelihood is binomial with 9 successes and one failure for true matches, and 8 successes and two failures for false matches. As it happens, the posterior distributions can be calculated analytically under these assumptions, and have, respectively, beta distributions with parameters (10,2) for the true matches and parameters (9,3) for false matches. These distributions are plotted in Figures 2 and 3.

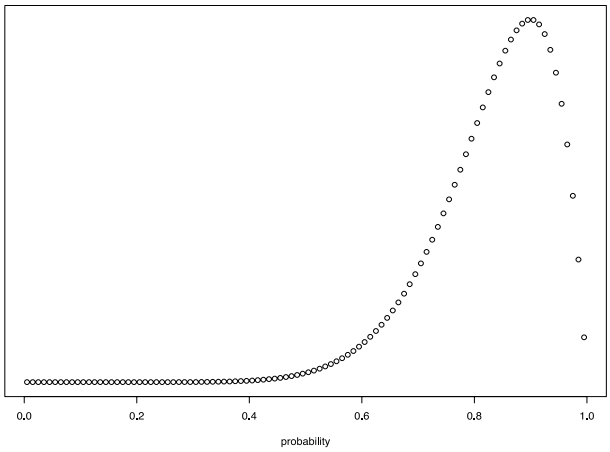


Figure 2: Posterior density for the probability of correctly assessing a true match.

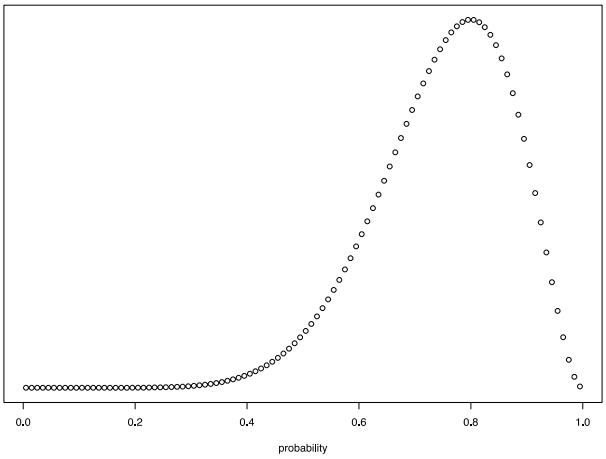


Figure 3: Posterior density for the probability of correctly assessing a false match.

In the Bayesian context, what happens to hypothesis tests and confidence intervals? When the posterior distribution is continuous, each sharp null hypothesis (such as p=1/2 in the example) has probability zero. Thus, if you ask a Bayesian for a probability for such a hypothesis, the answer can be given without reference to the data. Ask a better question! Intervals derived from the posterior distribution have the interpretation most people wish confidence intervals had – probabilities that the parameter lies in the specified interval.

It is no secret that Bayesian statistics is controversial, but less so now than a few decades ago. Some of the controversy revolves around the admittedly subjective nature of the prior distribution. Of course, the likelihood is subjective too. To put it most starkly, the data is the data, and everything else we make up. Of course, in order to retain our readers' interest, we have to explain why we made the choices we did. A careful analysis will highlight the assumptions to which the conclusions are most sensitive, so that readers can make their own judgment of how much to trust the conclusions.

The conversation between Bayesians and sampling theorists essentially revolves around what is considered to be random, and what is considered to be fixed. In Bayesian statistics, the data, once observed, are regarded as known and fixed at their observed values. The parameters, however, have distributions both before the data are observed (the prior) and after the data are observed (the posterior). By contrast, in sampling theory, the parameters are regarded as fixed but unknown, and the data are regarded as random even after the data are observed. There doesn't seem to be a way to compromise when the question is what is to be regarded as random and what is to be regarded as fixed.

4. Conclusion

This note shows how different styles of statistics apply to a hypothetical data set, and the basis for these different styles. The author finds himself attracted to the Bayesian viewpoint for the reasons above.

Joseph B. ("Jay") Kadane, Ph.D., is Leonard J. Savage University Professor of Statistics and Social Sciences, Emeritus, at Carnegie Mellon University. His research includes contributions to economics, political sciences, physics, chemistry, air pollution, internet security, philosophy, mathematics, marketing and law. He has experience as an expert witness in a variety of cases. His most recent book is "Principles of Uncertainty," published by Chapman and Hall in 2011. A companion volume, "Pragmatics of Uncertainty" is in preparation. kadane@stat.cmu.edu

References

1. Kadane J. Post-Mortem Root Banding, A Sceptical Review. Law, Probability and Risk, 2015. DOI 10.1093/pr/mgv002. http://lpr.oxfordjournals.org/content/early/2015/04/01/lpr.mgv002.
2. Dunn M, Kadane JB, Garrow J. Comparing Harm Done by Mobility and Class Absence: Missing Students and Missing Data. J Educ Behav Stat 2003; 28(3): 269-288.

Suggestions for further reading

Comparing methods of inference:
Barnett V. Comparative statistical inference. New York: Wiley and Sons, 1999.

Bayesian matters, easy introduction:
Lindley DV. Understanding Uncertainty. Hoboken, NJ: Wiley, 2006.
Lindley DV. Making Decisions. 2nd Edition, London: Wiley, 1991.

How to:
Gelman A, Carlin JB, Stern HA, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. 3rd ed., Boca Raton, FL: CRC Press, 2014.

Why to:
Kadane J. Principles of Uncertainty. Chapman and Hall. http://stat.cmu.edu/~kadane.

Section 1: Experimental Design and Statistics

Discussion Summary

The questions posed to the panel consisting of Constantine Gatsonis, Stephen L. Morgan, Dietram Scheufele, Marilyn London, and Joseph B. Kadane, ranged from issues with presenting Bayesian statistics in courts, to effectively establishing realistic studies within IRB regulations, to fundamental issues with the practice of forensic science. Kadane explained that as long as a scientist is clear on his or her basis of judgment, the specific type of statistical analysis should not be a problem in court. Scheufele acknowledged that there are many factors, such as the Hawthorne effect, which can lead to issues in behavioral studies; however, these can be overcome by intentional deception of subjects as long as that deception is debriefed at the conclusion of the experiment. He recommended looking at literature in the area of risk decision for examples of well-handled and legitimate studies. It was generally agreed that forensics poses a unique problem in that each case is a "study of one," as suggested by London, and that there is often no "ground truth" that can be used to validate results. Furthermore, the reliability of analyses can vary widely depending on the practitioner, and therefore pooled data may not be the most effective way to evaluate error rate. And while quantitation is the goal for most scientists, there are many different types of quantitation that must be statistically analyzed to determine their appropriateness in application.

Section 2: Interpretation and Assessment

Plenary II. Topic: Scientific Impact of Problematic Literature
Title: Pernicious, Pervasive, and Persistent Literature in Fire Investigation

John J. Lentini

As a forensic science discipline, fire investigation is unique in the amount of widespread, persistent and problematic literature affecting the beliefs and the behavior of its practitioners. The story begins in 1977, when Boudreau, Kwan and Faragher, working on an Aerospace Corporation grant from the Law Enforcement Assistance Administration (LEAA), conducted a "Survey and Assessment" of arson and arson investigation techniques. In that assessment, the authors listed seven "burn indicators," but stated, "Although burn indicators are widely used to establish the causes of fires, they have received little or no scientific testing." They recommended, "that a program of carefully planned scientific experiments be conducted to establish the reliability of currently used burn indicators," and "a handbook based on the results of the testing program should be prepared for field use by arson investigators."

Three years later, the "Handbook" called for in the survey and assessment was published by the most respected scientific and engineering body on the planet, the United States National Bureau of Standards (NBS). Unfortunately, the scientific studies recommended in the survey had not been conducted. The NBS editors, Brannigan, Bright, and Jason, were advised by two members of the National Fire Academy staff, and in Chapter 1 they repeated most of the myths that have been used to incorrectly determine that a fire burned faster or hotter than normal. The text refers to "hot" fires and a "rapid buildup of heat," which is generally interpreted by investigators as indicative of the use of liquid accelerants.

Given the imprimatur tour of such an august body, authors of fire investigation textbooks for the next 20 years felt perfectly comfortable publishing these myths. They were also cited in hundreds, if not thousands of reports where the investigators found arson even though the fires were accidental.

This paper will identify many of the circular references, and illustrate the damage that took so long to undo, although there are still many books in print, even new books, that cite the mythology.

Introduction

This article will explore the development and promulgation of the mythology of arson investigation through published literature. Certainly, there is no reason to believe that anyone ever set out to promulgate something that was not true. It is likely that many myths came about as a result of unwarranted generalizations. For example, an investigator might observe a pattern

of spalling around the remains of a gasoline container and make an association of gasoline with spalling. The next time that spalling is observed, gasoline is inferred.

Some myths arose because of intuitively "obvious deductions." The notion that gasoline burns hotter than wood is an appealing one, as is the notion that a narrow V-pattern indicates a "rapid fire." The problem is that the term "rapid" is never defined, thus making it impossible, in many cases, to actually design an experiment to test a particular hypothesis about the significance of a particular indicator. Even when an indicator can be shown by direct evidence to be of no value, resistance to change and a culture of "circular citations" allow the myth to live on. Most of these circular citations happen in fire investigation textbooks. The errors in the peer-reviewed literature, with some exceptions, occur with far less frequency.

Many of the myths were gathered by Boudreau, Kwan and Faragher, working for the Law Enforcement Assistance Administration (LEAA) and published in *Arson and Arson Investigation: Survey and Assessment* (1977).[1] Although the myths were reported with appropriate cautionary language, the cautions were not heeded. Three years later, when the "indicators" were listed by what should have been the ultimate authority, the cautions were lost. No less an authority than the National Bureau of Standards (NBS then—now NIST) published a *Fire Investigation Handbook* (1980), which stated that crazed glass meant rapid heating, shiny alligator blisters meant that a fire burned "faster than normal," and narrow V's indicate "fast-developing, hot fires."[2] Fire investigators invariably equated these descriptions of the fire as "accelerated using a flammable liquid."

In the 1980s, one American text after another referred to the NBS publication or to another publication that cited the myths published in the LEAA report. These circular citations continue in books still in print. Interestingly, many of the myths never gained much credibility in the United Kingdom because the major "go to" textbook, Cooke and Ide's *Principles of Fire Investigation* (1985), either did not repeat the myths, or provided an accurate interpretation of the significance of indicators such as crazing and spalling.[3]

NFPA 921

In 1985, when the National Fire Protection Association (NFPA) Standards Council became sufficiently concerned about the validity of fire investigations, it appointed a Technical Committee to address the issue. Seven years later, the Committee and NFPA produced the first edition of NFPA 921, *Guide for Fire and Explosion Investigations.* The document listed many of the myths from the LEAA study and the NBS Handbook, and stated why these "indicators" were "misconceptions." The howls of protest from fire investigation "professionals" were deafening. If what was printed in that document were actually true, it meant that hundreds if not thousands of accidental fires had been wrongly determined to be incendiary fires. No investigator wanted to admit to the unspeakable possibility that they had caused an innocent person to be wrongly convicted, or a family to be wrongly denied their life savings. The profession was in denial, and cited the older publications as support for that denial.

In 1998, the Technical Committee on Fire Investigations, responding to public pressure, removed the word "misconception" from the titles of several paragraphs in the chapter on pattern

development in the optimistic but mistaken belief that previous editions of the document, which was still not accepted in many quarters, had relieved the profession of these misconceptions.

The myths are slowly dying out (or being *"Dauberted"* out), but there are still practitioners who use them today, with disastrous consequences. Apparently, it is more difficult to "unlearn" a myth if one is not equipped with the kind of "baloney detector" that is, one hopes, instilled in students in undergraduate science programs. A recent survey of 217 investigators employed by fire departments revealed that, NFPA 921 notwithstanding, 33% of them believe that crazed glass indicates the presence of an accelerant.[4]

It is distressing to find mythology printed in law enforcement oriented textbooks, but what investigators do with the myths is much worse. Through confident expert testimony, misguided investigators persuade juries that their opinions are correct, and somebody set the fire in question. If such were true, there is often only one possible fire setter. And prior to 1995, it was uncommon for a criminal defendant to have an expert on his or her side to counter the bad science. The NBS Handbook stated, "In general, the function of the fire investigator will be to support prosecution by: …"[5]

Some examples are presented below.

MI vs. David Lee Gavitt

David Gavitt served 27 years for setting the fire that killed his wife and two children. The prosecutors could find no motive but proceeded based on a finding of "alligatoring" and a deeply flawed chemical analysis, which identified gasoline. Here is the "expert" testimony describing the charring:

> "When the gases come out, they dig little trenches for themselves, and all of a sudden the wood begins to look like the back of an alligator. We call this alligatoring. The hotter the fire, the deeper the trenches. Way down inside of the trench soon starts to cool, so that soot turns to charcoal, insulates the inner wood, can't pyrolyze it anymore."[6]

PA vs. Han Tak Lee

Daniel Aston, a part time fire investigator, who claimed to have investigated 15,000 fires in 20 years was allowed to render the following testimony:

> "A dull alligatoring indicates slow, very slow fire." Mr. Aston repeated his mistaken belief about what alligatoring indicated again at page 488 ("extremely hot fire, flammable liquid, combustible liquid, once again"), and once more at page 493 ("The severe alligatoring and checking of this char…indicating extremely low fire….flammable, combustible liquid, once again.")[7]

No flammable liquids were detected in the debris, despite Aston's precise calculation that more than 60 gallons were involved. Han Tak Lee served 25 years for setting this fire.

AZ vs. Ray Girdler

Ray Girdler served eight years for setting the fire that killed his wife and daughter. He became a suspect when the fire chief noted that he was fully dressed at 2 AM. The Chief did not

manage to learn that Ray had fled his home barefoot and clad only in undershorts, and had been given clothing by his neighbor. The Fire Marshal, David Dale, was able to find plenty of "arson indicators," including crazed glass. Dale's testimony, which was rife with other myths, included the following.

> Much of this glass was heavily heat-crazed and free of smoke deposit, indicating a rapid buildup of intense heat and close proximity to initial fire, which is consistent with the expected results of a liquid accelerant fire.[8]

PA vs. Paul Camiolo

Mr. Camiolo was held for 10 months in jail awaiting his capital murder trial for setting the fire that killed his elderly parents. One of the investigators for the Commonwealth in his civil rights case relied on crazed glass. George Wert wrote, "A photograph of melted "crazed" glass indicates a very rapid build up of heat in the family room. This indicated a very rapid spread of fire unlike a cigarette fire which would burn much more slowly."[9]

TX vs. Cameron Todd Willingham

In this, the most famous of all botched arson cases, the Fire Marshal wrote, "The pieces of broken window glass on the ledge of the north windows to the northeast bedroom disclosed a crazed (spiderwebbing) condition. This condition is an indication that the fire burned fast and hot."[10]

Too many to count

The author has participated in more than 50 cases in which uninformed investigators, usually certified fire investigators with many years of experience, testified, or were prepared to testify that arson had occurred, based on the myths they learned from textbooks. There is no way to determine how many citizens have been wrongly imprisoned or wrongly denied the proceeds of an insurance policy as a result.

If one examines the peer-reviewed literature, however, the myths will not be found. It is the near absence of peer review that allows texts to be sold that promulgate the errors. There does not seem to be any obvious solution to the problem. As late as 2013, the tenth edition of *Criminal Investigation*, a $200 textbook that forms the basis of the curriculum in many criminal justice classrooms, repeated the myths from the 1977 *Survey and Assessment*. Most of the fire investigation texts produced in this century have embraced better science, but Amazon.com still has many of the books from the 1990s available, and the publishers are only too happy to print extra copies if they sense a demand.

The situation is not helped by the fact that public sector salaries for fire investigators are insufficient to attract applicants who have had any kind of science education since high school. Many fire investigators in practice today are unable to name the basic units of energy, or to define a watt. Unless and until governments are willing to spend the money required to hire people who understand the fundamentals of fire science, fire origin and cause determinations will continue to suffer from high rates of error, and all that entails.

John Lentini, CFI, D-ABC, is a consultant for Scientific Fire Analysis, LLC. Mr. Lentini is one of a handful of people certified to conduct both fire scene investigations and fire debris analysis. He has personally conducted more than 2,000 fire scene inspections and has appeared as an expert witness on more than 200 occasions. He is a frequent invited speaker on fire investigation science, and an active proponent of standards for fire and other forensic investigations. He is a member of the NFPA Technical Committee on Fire Investigations, and has served three terms as chair of ASTM Committee E30 on Forensic Science. John is the current Chairman of the AAFS Criminalistics Section. He also serves on the NIST/OSAC Subcommittee on Fire and Explosion Investigations. He is now an independent consultant living in the Florida Keys and doing business as Scientific Fire Analysis. His book, *Scientific Protocols for Fire Investigation*, was published by CRC Press in 2013 and is now in its second edition. scientific.fire@yahoo.com, www.firescientist.com.

References

1. Boudreau J, Kwan Q, Faragher W, Denault G. Arson and Arson Investigation-Survey and Assessment. Washington, DC: Law Enforcement Assistance Administration; United States Department of Justice, 1977.
2. Brannigan F, Bright R, Jason N. Fire Investigation Handbook. Washington, DC: National Bureau of Standards, U.S. Department of Commerce, 1980.
3. Cooke R, Ide R. Principles of Fire Investigation. Leicester, UK: The Institution of Fire Engineers, 1985.
4. Cook D. Have We Learned the Lessons of the Willingham Case? A National Survey of Fire Investigators, La Jolla, CA: Master's thesis, National University, 2015.
5. Branigan et al. Appendix B, 1980: 175.
6. Michigan vs. David Lee Gavitt, (1985) Testimony of Dr. Ronald Edwards at page 13, lines 12-15, Ionia County, Case No. 85-007555.
7. Pennsylvania vs. Han Tak Lee, (1989) Testimony of Daniel Aston, P.E., Common Pleas Court of Monroe County, Excerpts of trial testimony are from the affidavit prepared by John Lentini, filed in December 2005.
8. Arizona vs. Ray Girdler, Jr., (1983), in the Superior Court of the State of Arizona in and for the County of Yavapai, No. 9809, Testimony of David Dale.
9. Wert, George, Report of September 5, 1997 addressed to Mark Miller, Montgomery Country District Attorney's Office, p 5.
10. Vasquez M. Report of January 24, 1992 on the Willingham fire, p 5.

Why Scientists Make Mistakes in Conducting and Reporting their Research

Michael Shermer

The evaluation of evidence and the formation and testing of hypotheses and the development of causal theories to explain heretofore unexplained phenomena is one of the primary topics of interest of *Skeptic* magazine and the "Skeptic" column in *Scientific American,* for which the author is editor and writer. As well his 1997 book, *Why People Believe Weird Things*, covers at length how the scientific process works, and his 2011 book, *The Believing Brain*, deals with the very serious problem of cognitive biases and how they distort our perception of the world and evaluation of claims and evidence.[1,2]

More specifically, many false claims that scientists failed to catch are related to forensic science and have landed people in jail and ruined their lives utterly. Such was the case of the Satanic Panic of the 1980s involving so-called "Satanic Ritual Abuse"[3,4,5] and the Recovered Memory Movement of the 1990s,[6,7] both of which caused irreparable harm to many people who were accused of serious crimes, from animal cruelty and ritual abuse to child molestation and even murder. Many were charged, tried, and convicted on nothing more than bogus evidence—junk science in the courtroom as it came to be called. And it wasn't just the misreading of a fingerprint, bite mark, or fire burn, but the unquestioned acceptance of unsubstantiated claims, such as that Satanic cults were proliferating American cities (a mutilated cat, for example, became evidence of such), or that preschools around the country—starting with the McMartin Preschool in Southern California—were dens of ritual abuse that involved secret tunnels (that no one could find), impossible trips to far off places (during school hours with people all around), and other ridiculous tales that later unraveled in court appeals.

The "recovered memory movement" is a case in point. It is based on the false premise that memories of childhood sexual abuse are repressed by the victims but recalled decades later through use of special therapeutic techniques, including suggestive questioning, hypnosis, hypnotic age-regression, visualization, sodium amytal ("truth serum") injections, and dream interpretation, none of which have any evidence in support of their validity as a memory retrieval method. It begins with a "psychotherapist" having a client read books about recovered memories, watch videotapes of talk shows on recovered memories, and participate in group counseling with other women how have recovered memories of abuse. Absent at the beginning of therapy, memories of childhood sexual abuse are soon created through weeks and months of applying these special therapeutic techniques. Then, as in the witch crazes of centuries past, names are named—father, grandfather, uncle, brother, friends of father, and so on. Next is confrontation with the accused, who inevitably denies the charges, and termination of all relations with the accused. Shattered families are the result. Experts on both sides of this issue estimate that at least one million people have "recovered" memories of sexual abuse since 1988 alone, and this does not count those who really were sexually abused and never forgot it.

This movement began with an idea—a bad idea—published in psychotherapist Jeffrey Masson's 1984 book, *The Assault on Truth*, in which he rejected Freud's claim that childhood sexual abuse was fantasy and argued that Freud's initial position—that the sexual abuse so often recounted by his patients was actual, rampant, and responsible for adult women's neuroses—was the correct one.[8] The movement became a full-blown witch craze when Ellen Bass and Laura Davis published *The Courage to Heal: A Guide for Women Survivors of Child Sexual Abuse* in 1988.[9] One of its conclusions was "If you think you were abused and your life shows the symptoms, then you were." The book sold more than 750,000 copies and triggered a recovered memory industry that involved dozens of similar books, talk-show programs, and magazine and newspaper stories.

But it was a witch craze, not a sex craze. The supposed numbers alone should make us skeptical. Bass and Davis and others estimate that as many as one-third to one-half of all women were sexually abused as children. Using the conservative percentage, this means that in America alone 42.9 million women were sexually abused. Since they have to be abused by someone, this means about 42.9 million men are sex offenders, bringing us to a total of 85.8 million Americans. Additionally, many of these cases allegedly involve mothers who consent and friends and relatives who participate. This would push the figure to over 100 million Americans (about 38 percent of the entire population) involved in sexual abuse. Impossible. This movement is made all the scarier by the fact that not only can anyone be accused, the consequences are extreme—incarceration. Many men and a number of women have been sent to jail, and some are still sitting there, after being convicted of sexual abuse on nothing more than a recovered memory.

How do such travesties happen? The answer is found in the brain, or more precisely how the brain processes information. Once we form beliefs and make commitments to them, we maintain and reinforce them through a number of powerful *cognitive heuristics* that guarantee they are correct (at least in our minds). A *heuristic* is a mental method of solving a problem through intuition, trial and error, or informal methods when there is no formal means or formula for solving it (and often even when there is). These heuristics are sometimes called *rules of thumb*, although they are better known as *cognitive biases* because they almost always distort percepts to fit preconceived concepts. Beliefs precede perceptions. There are a number of specific cognitive heuristics that operate to confirm our beliefs. Here are a few. I cover them in much more detail in Chapter 12 of *The Believing Brain*.

The Confirmation Bias

The confirmation bias is *the tendency to seek and find confirmatory evidence in support of already existing beliefs and ignore or reinterpret disconfirmatory evidence*. It is best captured in the biblical wisdom: *Seek and ye shall find.* Experimental examples abound.[10]
In 1981, the psychologist Mark Snyder tasked subjects to assess the personality of someone whom they were about to meet, but only after they reviewed a profile of the person. One group of subjects were given a profile of an introvert (shy, timid, quiet), while another group of subjects were given a profile of an extrovert (sociable, talkative, outgoing). When asked to make a personality assessment, those subjects who were told that the person would be an extrovert tended to ask questions that would lead to that conclusion; the introvert group did the same in the opposite direction.[11]

In a 1983 study, psychologists John Darley and Paul Gross showed subjects a video of a child taking a test. One group was told that the child was from a high socioeconomic class while the other group was told that the child was from a low socioeconomic class. The subjects were then asked to evaluate the academic abilities of the child based on the results of the test. Even though both groups of subjects were evaluating the exact same set of numbers, those who were told that the children they were evaluating were from a high socioeconomic class rate the child's abilities as above grade level, and those who thought that the kids were from a low socioeconomic class rated the kids as below grade level in ability.[12] This is a striking indictment of human reason but a testimony to the power of belief expectations.

The power of expectation was on stage in a 1989 study by the psychologists Bonnie Sherman and Ziva Kunda, who presented a group of subjects with evidence that contradicted a belief they held deeply, and with evidence that supported those same beliefs. The results showed that the subjects recognized the validity of the confirming evidence but were skeptical of the value of the disconfirming evidence.[13] In another 1989 study by the psychologist Deanna Kuhn, when children and young adults were exposed to evidence inconsistent with a theory they preferred, they failed to notice the contradictory evidence, or if they did acknowledge its existence, they tended to reinterpret it to favor their preconceived beliefs.[14] In a related study, Kuhn exposed subjects to an audio recording of an actual murder trial and discovered that instead of evaluating the evidence first and then coming to a conclusion, most subjects concocted a narrative in their mind about what happened, made a decision of guilty or innocence, then riffled through the evidence and picked out what most closely fit the story.[15]

The Hindsight Bias

The hindsight bias is *the tendency to reconstruct the past to fit with present knowledge*. Once an event has occurred, we look back and reconstruct how it happened, why it had to happen that way and not some other way, and why we should have seen it coming all along.[16] Such "Monday-morning quarterbacking" is literally evident on the Monday mornings following a weekend filled with football games. We all know what plays should have been called…after the outcome. Ditto the stock market and the endless parade of financial experts whose prognostications are quickly forgotten as they shift to postdiction analysis after the market closes. It's easy to "buy low, sell high" once you have perfect information, which is only available after the fact when it is too late.

The hindsight bias is particularly on prominent display after a major disaster, when everyone thinks that they know how and why it happened, and why our experts and leaders should have seen it coming. NASA engineers should have known that the O-ring on the Space Shuttle Challenger's solid rocket booster joints would fail in freezing temperatures leading to a massive explosion, or that a small foam strike on the leading edge of the wing of the Space Shuttle Columbia would result in its destruction upon re-entry. Such highly improbable and unpredictable events become not only probable but practically certain *after they happen*. The hand-wringing and finger-pointing by the members of NASA's investigative commissions tasked with determining the causes of the two Space Shuttle disasters were case studies in the hindsight bias. Had such certainty really existed before the fact then of course different actions would have been taken.

The Self-Justification Bias

The self-justification bias is *the tendency to rationalize decisions after the fact to convince ourselves that what we did was the best thing we could have done*. Once we make a decision about something in our lives we carefully screen subsequent data and filter out all contradictory information related to that decision, leaving only evidence in support of the choice we made. This bias applies to everything from career and job choices to mundane purchases. One of the practical benefits of self-justification is that no matter what decision we make—to take this or that job, to marry this or that person, to purchase this or that product—we will almost always be satisfied with the decision, even when the objective evidence is to the contrary.

This process of cherry picking the data happens at even the highest levels of expert assessment. The political scientist Philip Tetlock, for example, in his book *Expert Political Judgment*, reviews the evidence for the ability of professional experts in politics and economics and finds that even though all of them claim to have data in support of their predictions and assessments, it turns out that when analyzed after the fact such expert opinions and predictions turn out to be no better than those of nonexperts—or even chance. Yet, as the self-justification heuristic would predict, experts are significantly less likely to admit that they are wrong than nonexperts.[17] Or as I like to say, *smart people believe weird things because they are better at rationalizing their beliefs that they hold for non-smart reasons*.

A very disturbing real-world example of the self-justification heuristic can be seen in the criminal justice system. According to the Northwestern University Law Professor Rob Warden, "You get in the system and you become very cynical. People are lying to you all over the place. Then you develop a theory of the crime, and it leads to what we call tunnel vision. Years later overwhelming evidence comes out that the guy was innocent. And you're sitting there thinking, 'Wait a minute. Either this overwhelming evidence is wrong or I was wrong—and I couldn't have been wrong because I'm a good guy.' That's a psychological phenomenon I have seen over and over."[18]

Anchoring Bias

Lacking some objective standard to evaluate beliefs and decision—which is usually not available—we grasp for any standard on hand, no matter how seemingly subjective. Such standards are called anchors, and this creates the *anchoring effect*, or *the tendency to rely too heavily on a past reference or on one piece of information when making decisions*. The comparison anchor can even be entirely arbitrary. In one study subjects were asked to give the last four digits of their Social Security numbers, and then asked to estimate the number of physicians in New York City. Bizarrely, people with higher Social Security numbers tended to give higher estimates for the number of docs in Manhattan. In a related study, subjects were shown an array of items to purchase—a bottle of wine, a cordless keyboard computer, a video game—and were then told that the price of the items was equal to the last two digits of their Social Security numbers. When subsequently asked the maximum price they would be willing to pay, subjects with high Social Security numbers consistently said that they would be willing to pay more than those with low numbers.[19] With no objective anchor for comparison, this random anchor made people more vulnerable to such arbitrary influence.

Availability Heuristic

Have you ever noticed how many red lights you encounter while driving when you are late for an appointment? Me too. How does the universe know that I left late and is punishing me for it? It doesn't, of course, but the fact that most of us notice more red lights when we are running late is an example of the *availability heuristic*, or the tendency to assign probabilities of potential outcomes based on examples that are immediately available to us, especially those which are vivid, unusual, or emotionally charged, which are then generalized into conclusions upon which choices are based.[20]

For example, your estimation of the probability of dying in a plane crash (or lightning strike, shark attack, terrorist attack, etc.) will be directly related to the availability of just such an event in your world, especially your exposure to it in mass media. If newspapers and especially television cover an event there is a good chance that people will overestimate the probability of that event happening.[21] An Emory University study, for example, revealed that the leading cause of death in men—heart disease—received the same amount of media coverage as the 11th-ranked vector: homicide. By contrast, drug use—the lowest ranking risk factor associated with serious illness and death—received as much attention as the second-ranked risk factor of poor diet and lack of exercise. Other studies have found that women in their 40s believe they have a 1 in 10 chance of dying from breast cancer, while their real lifetime odds are more like 1 in 250. This effect is directly related to the number of news stories about breast cancer.[22]

Representative Bias

Related to the availability bias is the representative bias, which is described by its discoverers, the psychologists Amos Tversky and Daniel Kahneman as: "an event is judged probable to the extent that it represents the essential features of its parent population or generating process." And, more generally, "when faced with the difficult task of judging probability or frequency, people employ a limited number of heuristics which reduce these judgments to simpler ones."[23] The following thought experiment has become a classic in cognitive studies: Imagine that you are looking to hire someone for your company and you are considering for employment the following candidate:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more likely? 1. Linda is a bank teller. 2. Linda is a bank teller and is active in the feminist movement.

When this scenario was presented to subjects, 85 percent chose the second option. Mathematically speaking, this is the wrong choice, because the probability of two events occurring together will always be less than the probability of either one occurring alone. And yet most people get this problem wrong because they fall victim to the representative fallacy, in which the descriptive terms presented in the second option seem more representative of the description of Linda.[24]

Hundreds of experiments reveal time and again that people make snap decisions under high levels of uncertainty, and they do so by employing these various rules of thumb to shortcut the computational process. For example, in another experiment policy experts were asked to estimate the probability that the Soviet Union would invade Poland and that the United States would then break off diplomatic relations. Subjects gave this a probability of 4 percent. Meanwhile, another group of policy experts was asked to estimate the probability just that the United States would break off diplomatic relations with the Soviet Union. Contrary to what the odds actually would be, these experts gave this latter scenario only a 1 percent chance of happening. The experimenters concluded that the more detailed two-part scenario seemed more representative of the actors involved.

The Blind Spot Bias

This bias is really a meta-bias in that it is grounded in all the other cognitive biases. The *bias blind spot* is *the tendency to recognize the power of cognitive biases in other people but to be blind to their influence upon our own beliefs*. In one study conducted by Princeton University psychologist Emily Pronin and her colleagues, subjects were randomly assigned high or low scores on a "social intelligence" test. Unsurprisingly, those given the high marks rated the test fairer and more useful than those receiving low marks. When asked if it was possible that they had been influenced by the score on the test, subjects responded that other participants had been far more biased than they were. Even when subjects admit to having such a bias as being a member of a partisan group, says Pronin, this "is apt to be accompanied by the insistence that, in their own case, this status…has been uniquely *enlightening*—indeed, that it is the *lack* of such enlightenment that is making those on the other side of the issue take their misguided position." In a related study at Stanford University, students were asked to compare themselves to their peers on such personal qualities as friendliness and selfishness. Predictably, they rated themselves higher. Yet, even when the subjects were warned about the *better-than-average bias* and asked to re-evaluate their original assessments, 63 percent claimed that their initial evaluations were objective, and 13 percent even claimed to be too modest![25]

* * *

The bottom line is that we should be skeptical of all claims, including and especially those that we really want or expect to be true. This includes scientists, the very people whom one might think would be immune to such biases, but in point of fact the more intelligent and educated a person is the better they are at rationalizing beliefs that they arrived at for non-rational reasons. That is to say, as some of the smartest and most educated people in the world scientists need to be especially cautious about not fooling themselves, as the Nobel laureate physicist Richard Feynman told the 1874 graduating class of Caltech: "The first principle is that you must not fool yourself—and you are the easiest person to fool."[26]

This problem of bias in science goes a long way toward explaining the findings of John Ioannidis in his widely read and controversial essay in *PLoS Medicine*, "Why Most Published Research Findings Are False." After noting a number of factors that increase the probability that a research claim or finding is false, including "when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical

modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance," Ioannidis concluded with this disturbing finding: "for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias." The solution? "Diminishing bias through enhanced research standards and curtailing of prejudices" may help, he notes, but Ioannidis also notes what in this author's opinion is the best work-around of all, a convergence of evidence from multiple sources: "most research questions are addressed by many teams, and it is misleading to emphasize the statistically significant findings of any single team. What matters is the totality of the evidence."[27]

This is what the nineteenth century philosopher of science William Whewell called a *consilience of inductions*. Whewell reasoned that in order to prove a theory one should have more than one induction, or a single generalization drawn from specific facts. One must have multiple inductions that converge upon one another, independently but in conjunction. Whewell said that if these inductions "jump together" it strengthens the plausibility of a theory being true: "Accordingly the cases in which inductions from classes of facts altogether different have thus jumped together, belong only to the best established theories which the history of science contains. And, as I shall have occasion to refer to this particular feature in their evidence, I will take the liberty of describing it by a particular phrase; and will term it the Consilience of Inductions."[28] Call it the convergence of evidence.

Just as detectives employ the convergence of evidence technique to deduce who most likely committed a crime, scientists employ the method to deduce the likeliest explanation for a particular phenomenon. Cosmologists reconstruct the history of the universe through a convergence of evidence from astronomy, astrophysics, planetary geology, and physics. Geologists reconstruct the history of the Earth through a convergence of evidence from geology, geophysics, and geochemistry. Archeologists piece together the history of civilization through a convergence of evidence from pollen grains, kitchen middens, potshards, tools, works of art, written sources, and other site-specific artifacts. As a historical science, evolution is confirmed by the fact that so many independent lines of evidence converge to this single conclusion. Independent sets of data from geology, paleontology, botany, zoology, herpetology, entomology, biogeography, comparative anatomy and physiology, genetics and population genetics, and many other sciences each point to the conclusion that life evolved. This is classic consilience of inductions.

The irony of Whewell's philosophy of science—and a telling one for the problem we are here addressing of bias in science—is that the theory of evolution is arguably the most consilient theory ever generated, and yet Whewell rejected it, going so far as to block the book from being shelved at the library at Trinity College, Cambridge. Some biases are just too powerful to overcome. Let that be a lesson for us all.

Michael Shermer, Ph.D., is the Founding Publisher of *Skeptic* magazine (www.skeptic.com), the Executive Director of the Skeptics Society, a monthly columnist for *Scientific American* and the host of the Skeptics Distinguished Science Lecture Series at Caltech. Dr. Shermer received his B.A. in psychology from Pepperdine University, M.A. in experimental psychology from California State University, Fullerton, and his Ph.D. in the history of science from Claremont

Graduate University (1991). He teaches a transdisciplinary course for Ph.D. students at Claremont Graduate University entitled "Evolution, Economics, and the Brain," and an honors course for undergraduates at Chapman University. He has been a college professor since 1979, also teaching psychology, evolution, and the history of science at Occidental College (1989-1998), California State University Los Angeles, and Glendale College. As a public intellectual he regularly contributes Opinion Editorials, book reviews, and essays to the Wall Street Journal, the Los Angeles Times, Science, Nature, and other publications. He has appeared on such shows as The Colbert Report, 20/20, Dateline, Charlie Rose, Larry King Live, Tom Snyder, Donahue, Oprah, Lezza, Unsolved Mysteries and other shows as a skeptic of weird and extraordinary claims, as well as interviews in countless science and history documentaries aired on PBS, A&E, Discovery, The History Channel, The Science Channel, and The Learning Channel. Dr. Shermer was the co-host and co-producer of the 13-hour Family Channel television series, *Exploring the Unknown*. mshermer@skeptic.com

References

1. Shermer M. Why People Believe Weird Things. New York: W.H. Freeman, 1997.
2. Shermer M. The Believing Brain. New York: Henry Holt, 2011.
3. Goode E, Nachman B-Y. Moral Panics: The Social Construction of Deviance. Cambridge, MA: Wiley-Blackwell, 1994.
4. Spanos NP. Multiple Identities and False Memories: A Sociocognitive Perspective. Washington, DC: American Psychological Association, 1996.
5. Victor J. Construction of Satanic Ritual Abuse and the Creation of False Memories. In: Believed-in-Imaginings: The Narrative Construction of Reality. DeRivera J, Sarbin T, editors. Washington, DC: American Psychological Association, 1998.
6. Freyd J. Betrayal Trauma: The Logic of Forgetting Childhood Abuse. Cambridge, MA: Harvard University Press, 1998.
7. Loftus E. The Reality of Repressed Memories. Am Psychol 1993; 48: 518-537.
8. Masson J. The Assault on Truth: Freud's Suppression of the Seduction Theory. New York: Farrar, Straus & Giroux, 1998.
9. Bass E, Davis L. The Courage to Heal: A Guide for Women Survivors of Child Sexual Abuse. New York: Reed Consumer Books, 1988.
10. Nickerson R. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. Rev Gen Psychol 1998; 2: 175-220.
11. Snyder M. Seek and Ye Shall Find: Testing Hypotheses About Other People. In: Social Cognition: The Ontario Symposium on Personality and Social Psychology. Higgins ET, Heiman CP, Zanna MP, editors. Hillsdale, NJ: Erlbaum, 1981; 277-303.
12. Darley JM, Gross PH. A Hypothesis-Confirming Bias in Labeling Effects. J Pers Soc Psychol 1983; *44*: 20-33.
13. Sherman B, Kunda Z. Motivated Evaluation of Scientific Evidence. Paper presented at the annual meeting of the American Psychological Society, Arlington, VA, 1989.
14. Kuhn D. Children and Adults as Intuitive Scientists. Psychol Rev 1989; 96: 674-689.
15. Kuhn D, Weinstock M, Flaton R. How Well Do Jurors Reason? Competence Dimensions of Individual Variation in a Juror Reasoning Task. Psychol Sci 1994; 5: 289-296.
16. Tetlock P. Expert Political Judgment: How Good is It? How Can We Know? Princeton: Princeton University Press, 2005. See also: http://www.innocenceproject.org/.
17. Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus & Giroux, 2013.

18. Tversky A, Kahneman D. Availability: A Heuristic for Judging Frequency and Probability. Cognitive Psychol 1973; 5: 207-232.
19. Combs B, Slovic P. Newspaper Coverage of Causes of Death. Journalism Quart 1979; 56: 837-843.
20. Glassner B. The Culture of Fear: Why Americans Are Afraid of the Wrong Things. New York: Basic Books, 1999.
21. Tversky A, Kahneman D. Availability: A Heuristic for Judging Frequency and Probability. In: Judgment Under Uncertainty. Kahneman D, Slovic P, Tversky A. Cambridge: Cambridge University Press, 1982;163.
22. Tversky A, Kahneman D. Extension versus Intuititve Reasoning: The Conjunction Fallacy in Probability Judgment. Psychol Rev 1983; 90: 293-315.
23. Fischhoff B. For Those Condemned to Study the Past: Heuristics and Biases in Hindsight. In: Judgment Under Uncertainty. Kahneman D, Slovic P, Tversky A, editors. Cambridge: Cambridge University Press, 1982; 335-351.
24. Pronin E, Lin DY, Ross L. The Bias Blind Spot: Perceptions of Bias in Self versus Others. Pers Soc Psychol B 2002; 28: 369-381.
25. Feynman R. Cargo Cult Science. In: Surely You're Joking Mr. Feynman. Feynman R, Leighton R. New York: W. W. Norton, 1985; 338-346.
26. Ioannidis J. Why Most Published Research Findings are False. PLoS Med 2005; 2: 124.
27. Whewell W. The Philosophy of the Inductive Sciences. London: J. W. Parker, 1840.

Re-Examining Peer Review

Orla M. Smith

A crucial goal of scholarly scientific journals is to publish high quality, robust and reproducible research results, and the process by which journals accomplish this goal is rigorous peer review. The key elements of peer review are selection, quality control and ensuring adherence to standards. Confirming the quality of published work is becoming increasingly difficult as a result of data proliferation and the highly interdisciplinary nature of many research studies. Nevertheless, recent concerns from academia, industry, government and journals themselves over the problem of data irreproducibility highlight the enduring need for rigorous peer review. Against this backdrop, the traditional process of peer review has come under scrutiny and has been criticized. Thus, other models of peer review---pre- and post-publication peer review and peer review through crowdsourcing---have been proposed to complement or perhaps even replace the current process.

Let's take a closer look at the peer review process. The *Science* family of journals uses a rigorous, two-step peer review process. The majority of papers are evaluated by one or more members of an extensive Board of Reviewing Editors, composed of working scientists, who provide informal feedback to the editorial team about the paper's novelty and suitability for the journal. They do not perform in-depth analysis of the paper. Taking into account this feedback, the editorial team makes a decision to either send the paper for full peer review or to reject the paper, if it is unlikely to be suitable for the journal. At *Science Translational Medicine*, about 30% of papers are sent for full in-depth peer review to at least two and usually three expert reviewers. Selection of knowledgeable, reliable reviewers is essential to ensure a rigorous and fair peer review process.

So what are the key elements of a useful review? A good review provides a short synopsis of the paper, evaluates the quality of experiments and the validity of the interpretations, comments on the appropriateness of controls, and may suggest further experiments. A good review also discusses the impact the paper might have in its own field and, in a broader context, in other fields and notes any similar publications that might undercut the novelty of the presented results. About 80-85% of papers reviewed at *Science Translational Medicine* are rejected after either initial screening by the Board of Reviewing Editors or after one round of in-depth peer review. For the remainder, a second round of in-depth peer review (usually by the same reviewers) may be required if the revised paper incorporates new data or analyses requested by the reviewers. Only rarely will a paper be sent for more than two rounds of review.

Although traditional peer review is still the accepted measure by which scientific rigor and excellence are assessed, it has come under increasing scrutiny and criticism. The principal issues are that traditional peer review is too slow, is not fair, results in too many unnecessary rounds of review, lacks transparency, does not provide credit to the hard-working reviewers who provide their time and expertise for free, leads to redundant peer review at multiple journals, is of uneven quality, and does not provide a forum for post-publication discussion. Journals have

taken steps to try and improve the peer review process. For example, over the past year *Science* has introduced a process called cross-review. During cross-review, which occurs before the authors are allowed to see the reviewer comments on their paper, the reviewers have the opportunity to see and comment on the reports of the other reviewers and also to amend their own review if they wish. The goal of cross-review is to make the peer review process more uniform and more transparent. In another recent effort to improve peer review transparency, the American Association for the Advancement of Science (AAAS), the publisher of the *Science* family of journals, has acquired a Peer Review Evaluation service (PRE) that will be rolled out to AAAS journals in late 2015. This web-based service provides a badge of verification for each peer-reviewed research article and displays information regarding each step of the process (e.g. whether the paper was evaluated by a statistical reviewer, how many reviewers provided in-depth reviews of the paper).

Science Translational Medicine has experimented with a unique form of pre-review through crowdsourcing in collaboration with the Sage Bionetworks DREAM Breast Cancer Prognosis Challenge.[1] A computational challenge was presented to the scientific community: Devise an algorithm to predict, more accurately than current assays, the prognosis of breast cancer patients from clinical information (age, tumor size, histological grade, 10-year survival data) and gene and copy number expression data. Computational models were submitted as re-runnable source code to an open-access computational platform where they could be independently validated by other groups. Computational models were successively improved by such iterative testing, and the winner of the challenge was determined by subjecting the top models to further validation with another breast cancer patient dataset. Peer reviewers, selected by the *Science Translational Medicine* editorial team, were embedded throughout the challenge process as part of the organizing committee. The winning computational model was published in *Science Translational Medicine* with input from our Board of Reviewing Editors but without traditional peer review because in this case the model had already been thoroughly validated and peer reviewers had commented at each step of the challenge process, substituting for standard in-depth peer review.

In addition to the efforts of journals to improve traditional peer review, a number of new online initiatives have sprung up with the goal of addressing the imperfections of peer review in different ways. For example, Peerage of Science offers a free prepublication peer review process.[2] Authors can post any unpublished research paper, and any scientist (with at least one published paper) can provide comments, usually within 2 weeks. Reviewers and their reviews are rated as a way to ensure quality and to allow for academic credit. There is a standardized review format and an index of article quality. Journals that subscribe to Peerage of Science can track papers of interest and their associated reviews and can offer to publish articles. Authors can export the reviews to the website of any subscribing journal. An alternative to Peerage of Science is Rubriq, an independent, fee-based prepublication peer review process run by 100 published postgraduates.[3] The goal of Rubriq is to avoid redundant reviewing and to provide a rapid review process. Comments are obtained from three independent reviewers within two weeks and are presented on a standard score card, which provides an index of quality called the R-score. The comments are combined into a full evaluation of all aspects of the paper called the Rubriq Report. The author pays $600 for this service.

Open Scholar C.I.C, composed of a community of scholars, librarians and editors, now offers LIBRE, a platform for author-guided open peer review that uses open metrics and crowdsourcing to evaluate research papers at any step before, during or after peer review.[4] Authors invite peers to evaluate their posted work. Reviewers submit a signed, detailed, quantitative and qualitative review. The reviews are posted and they themselves are evaluated. Another post-publication peer review option is PubPeer, which hosts comments on any research article published in PubMed or with a DOI (Digital Object Identifier) and notifies authors when comments are posted.[5] All comments are consolidated into a central and searchable online database, providing an online discussion forum for the scientific community. Those posting comments can reveal their names if they wish, but the majority of comments are anonymous.

In summary, traditional peer review is still the gold standard for assessing the excellence, scientific rigor and novelty of research articles. However, many new online options are available that can work in concert with traditional peer review to improve the quality and transparency of this process with benefits for all.

Orla M. Smith, Ph.D., is Managing Editor of *Science Translational Medicine*, published by the American Association for the Advancement of Science (AAAS). Dr. Smith has a Ph.D. in Biochemistry from the Royal Free Hospital School of Medicine, University of London, and did postdoctoral work on the cell and molecular biology of stem cells at the Johns Hopkins Medical Institutions in Baltimore, Maryland. She comes to this position from the journal Cell, where she was the founding Editor of the Leading Edge section with responsibility for all front-end content, the popular SnapShot format and the Cell podcast. Prior to her time at Cell, Dr. Smith was Biology Perspectives Editor at Science where she also handled and edited manuscripts on neurodegenerative disease research. She began her career in scientific publishing as News and Views Editor at the journal Nature Medicine.

References

1. Cheng W-Y, Yang T-H O, Anastassiou D. Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment. Sci Translational Med 2013; 5(181): 181ra50.
2. Peerage of Science. www.peerageofscience.org
3. Rubriq. www.rubriq.com
4. LIBRE. www.openscholar.org.uk/libre
5. PubPeer. https://pubpeer.com

Recognition and Mitigation of Cognitive Bias in Forensic Science: From Crime Scene Investigation to Forensic Research and Literature

Itiel Dror

Introduction

Forensic science, like most other sciences and human endeavors, involves human perception, interpretation, judgement and decision making. While other domains have acknowledged the role of the human and have taken steps to mitigate and minimize vulnerabilities (such as cognitive bias), forensic science for the most part has denied its relevance and importance. For many decades forensic science claimed to be 'objective' and practically 'infallible.'[1] From a cognitive perspective (which I come from) it is quite astonishing to hear such claims, but understanding that forensic scientists often work within the adversarial legal system and that errors are not apparent sheds some light on this.

In the past years the forensic science domain has started to undergo a substantial transformation.[2] With growing research and attention to the role of the human examiner, it is becoming very clear that bias and other cognitive issues are at the heart of forensic work.[3] In addition to the cognitive research and literature which equivalently shows the effects of bias,[4] the findings that human confirmation bias has contributed to forensic error[5] and the National Academy of Sciences[6] report have all further contributed to the transformation in the forensic sciences.

With forensic scientists starting to receive education and training on cognitive issues, their growing understanding of the role that context and other cognitive factors have on human expert performance, are paving the way for further improvements in forensic science.

Bias in crime scene investigations

`The naïve view of crime scene investigators (CSIs) is that they merely gather evidence from the crime scene for the forensic laboratory to examine. However, many decisions actually take place at the crime scene.[3] CSIs have to determine where to look for evidence, whether what they find is actual data or noise, and then to decide whether evidence that they do find has sufficient information to justify sending it to the laboratory. All of these are more than mere technical collection of evidence, they are important decisions made at the crime scene, critical decisions, because they determine what evidence is collected, and what is not collected often is gone and cannot be retrieved later. Furthermore, CSIs also make clear evaluative decisions beyond those relevant to evidence collection. For example, they examine and determine bullet trajectories, analyze blood spatter patterns, and sometimes decide whether a crime has even been committed.

The CSI work requires context so as to guide them in their work. However, contextual information can bias how they collect evidence, where they look for it, and their entire approach to a crime scene. For example, a detective briefing a CSI that they believe that the burglary is not real but an attempt at insurance fraud, or that the dead person did not really commit suicide but

was murdered, affects how the CSI conducts their work. This is based on briefing, beliefs, that drive the CSI work prior to even arriving at the crime scene –i.e., it is not based on the actual crime scene. Similarly, investigating forensic evidence at a suspicious fire is biased if the examiner is first briefed that the owner of the property is in debt, has been unsuccessfully trying to sell the property and that they are going to get a very good payment from the fire insurance company. All of these, prior to seeing the burned property, bias the perception, interpretation and judgement of the investigator.

Bias in the forensic laboratories

When evidence arrives at the forensic laboratory, before it is even used for identification, it first has to be perceived. For example, a latent fingerprint mark consists of minutia (characteristics in the friction ridge flow, such as a bifurcation when it divides and separates into two). However, such analysis, the initial perception of the evidence is highly subjective and influence by a variety of cognitive factors.[7] This is demonstrated by the variability among forensic examiners (Table 1).

	LATENT FINGERPRINT									
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
	22	9	15	8	9	3	8	11	7	10
	21	11	25	7	10	9	9	10	6	5
	19	9	18	10	7	9	15	19	6	6
	21	21	29	14	12	9	8	9	4	8
	17	16	15	11	16	9	7	12	5	5
	20	14	22	9	10	7	13	18	7	9
	22	17	15	10	10	8	11	24	8	11
	9	9	19	6	9	8	18	16	9	10
	30	15	25	10	12	12	19	22	12	17
	25	13	18	13	12	10	13	15	7	10
SD	**5.49**	**4.01**	**4.93**	**2.49**	**2.45**	**2.32**	**4.25**	**5.15**	**2.23**	**3.54**

Table 1: The number of minutia observed by different latent print examiners analyzing the same fingerprint.[7]

The problem of variability, lack of reliability, and the influence of cognitive factors is even more apparent and pronounced when observing the variablity not across forensic examiners (whether different examiners observe the same number of minutia in the same fingerprint, i.e.,

are examiners consistent with one another) but when observing reliability within the same examiner at different times (whether the same examiner, examining the same prints, will be consistent with themselves). The results are presented in Table 2.

	LATENT FINGERPRINT									
	A	B	C	D	E	F	G	H	I	J
	1	1	4	1	1	2	3	2	0	1
	8	3	5	1	1	2	2	5	2	2
	1	3	3	3	6	4	9	9	1	2
	2	3	2	5	0	1	1	0	0	1
	6	2	2	3	4	1	3	3	0	3
	9	4	2	1	4	6	0	5	1	1
	0	4	5	2	4	3	3	7	0	0
	3	1	4	0	6	2	1	4	2	0
	4	3	9	0	4	4	3	1	1	3
	1	0	0	1	4	1	4	1	0	0
MEAN	**3.5**	**2.4**	**3.6**	**1.7**	**3.4**	**2.6**	**2.9**	**3.7**	**0.7**	**1.3**

Table 2: The difference in the number of minutia observed by the same latent print examiner looking at the same fingerprint at different times (the table presents the number of minutia observed at time 1 minus the number of minutia observed at time 2, i.e., zero indicates consistency[7]).

The initial perception of the evidence is further influenced by other cognitive factors, such as the presence of another fingerprint for comparison. This is why it is critical to adopt procedures such as Linear Sequential Unmasking (LSU) which minimize cognitive biases.[8]

These perceptual issues have been demonstrated in a number of forensic domains, including the more structured domains of fingerprinting and DNA, and we can confidently assume that they apply equally well (if not more) to many other forensic domains.

Furthermore, the cognitive factors and bias in the initial stage of perception of the evidence have more impact and power when considering the comparative stage when evidence is compared against a target suspect to determine their relative similarity and conclude their possible match, or even with the use of statistical tools. These have been shown to be

inconsistent and potentially biased, i.e., different examiners can reach different conclusions or different statistical results.[9-12]

Cognitive bias issues in the forensic laboratories also arise from the verification processes.[3] Issues around verification pertain to base-rate biases, where identifications are most often verified, thus causing an a priori expectation to verify (such cognitive issues also pertain to use of technology and database searches[13]). Other issues pertain to the verification process not being blind, e.g., knowing who made the initial decision, what they based their decision on, etc. In fact, in many forensic laboratories, the examiners choose who will verify their work. There are other issues concerning what happens when an identification is not verified.

All of these put the human examiner in the forensic laboratory in a cognitively problematic position, making it hard for them to achieve an independence of mind and make determinations based solely on the evidence. Other factors affecting and biasing the work in the forensic laboratory have to do with the exposure of examiners to irrelevant contextual information (e.g., whether the suspect confessed to the crime), working in an adversarial culture (see below). The different sources of biasing factors are summarized in Figure 1.

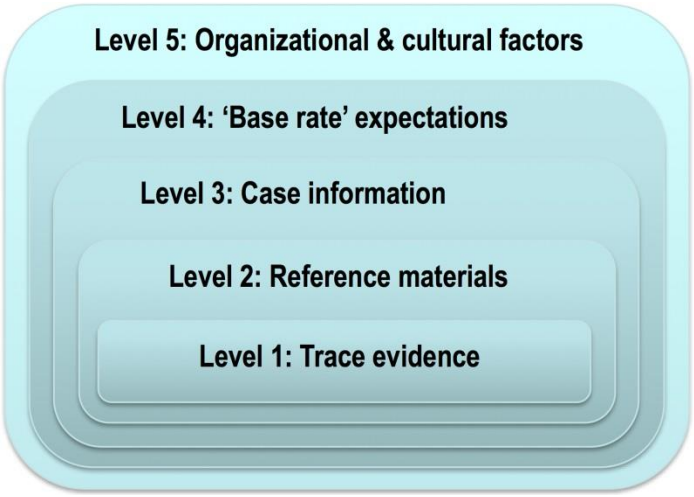


Figure 1: Different potential sources of bias.[3]

Bias in the court

Forensic scientists often work within the adversarial legal system, which is very problematic from a scientific point of view. First, forensic examiners are often brought into the legal proceedings to strengthen a case (for, or against a suspect). Hence, they are often recruited to play a role that is not scientific, that is, the opposing sides have little interest in the evidence per se, but more the need to have the forensic examiners provide testimony to strengthen their case (even during the investigation stage, forensic examiners are sometimes brought in to help support the existing beliefs of the investigating team).

Second, although the courts want to use science, the courtroom is not the place to do science (they do not have the time or the expertise). And what often happens is that rather than the courts using science, science is misused (and even abused) in the courtroom, all within the motivation of the different sides in the adversarial legal system.

Third, as a result of the adversarial legal system the forensic scientists often overstate their conclusions and confidence, and do not point out the limits and uncertainties of their conclusions (this is especially critical when many forensic domains do not have clear decision criteria –e.g., when 'two patterns are *similar enough* to conclude that they come form the same source', be it handwriting, bullet cartridges, fingerprints, or CCTV images).

Fourth, the forensic examiners many times present themselves (and they actually believe it) as objective and impartial, and not affected by irrelevant case information and other irrelevant contextual information that they have been exposed to. Fifth, nevertheless, the courts for the most part have accepted the unsubstantiated and unsupported claims made by the forensic scientists.

These have all furthered the problem of bias within forensic science.

Bias in forensic science research and literature

Three main factors are at play that have produced bias in forensic research and literature. First, forensic science exists and operates within a non-scientific environment –what is probably even an anti-scientific adversarial legal system (as I have elaborated above). Forensic science research and literature is often used in court, and that has impacted on what research is conducted, how it is conducted, and even if it is published.

Second, much of the research and literature in forensic science has (and is) been produced retroactively to justify existing practices. Rather than doing the research and then devising practices based on research and data, what has transpired (and is transpiring) is the exact opposite. Forensic practices (for the most part) have developed basically without any research or data, these have been accepted by the courts, and have been practiced for decades. Then with the growing scrutiny and criticism of forensic practices (see, e.g., the NAS Report[6]), the forensic community has been working on producing and finding the research to support the existing practices.

Thus, much of the research in forensic science has been motivated to underpin and justify the existing practices. This is, obviously, highly biased. This is further exacerbated by much of this research being led and conducted by the forensic examiners themselves, without proper input and guidance from non-interested parties. Furthermore, much of the literature and research is published (and reviewed) within the forensic community itself. Beyond the forensic community's perceived need and motivation to justify and underpin their existing practices (as well as defend itself from the criticism), the consequences of research findings that show major flaws in forensic work bias the way research is conducted, findings are interrupted, and the conclusions reached.

This is a major problem. Forensic research needs to be conducted 'context free', that is, without the motivation to defend existing practices and fear from findings that show flaws in how forensic science is currently practiced and used in court. The literature produced by forensic examiners to defend their ways has produced biased research and literature.

Third, there is a basic lack of a scientific culture in forensic science.[14] In addition to the issues already stipulated above, there are further issues that pertain to the forensic community being open and transparent about what they do. For forensic research and literature to be done properly forensic work needs to be documented, procedures and protocols should not be laboratory secrets, they (along with data) should be shared, and errors should be used for learning and improvement. These changes are not easy to make, but they (along with acknowledging limitations and uncertainties, and taking steps to minimize bias) are important in establishing a stronger scientific culture in forensic science. Without a research culture much of the research and literature will suffer from bias.

Conclusions

After of decades of forensic evidence been used in court, much of it without proper scientific foundation or validation, and much of it further weakened by cognitive bias, forensic science has stated to go through a transformation. The role of the human examiner and their susceptibility to cognitive bias, and the effects of irrelevant contextual information are current issues that the forensic community is working to address.

With new policies and training addressing cognitive issues in forensic science, the quality of work and the contribution of this domain will only increase.[2,15,16] However, we need to acknowledge the biasing factors and pressures of the adversarial legal system and the perceived need to justify existing practices, so we can take measures to minimize such influences. The literature and research underpin these advances and we need to make sure that these are not biased by interested parties and are not published and scrutinized by 'in house' similar minded parties.

Itiel Dror, Ph.D. is a cognitive neuroscientist at the Centre for the Forensic Sciences, University College London. He received his Ph.D. from Harvard in the area of cognitive factors in human expert performance. His insights and understanding of the human brain and cognitive system underpin his applied work with the US Air Force, medical experts, as well as forensic science and other expert domains. In the forensic field he has led to the understanding that the perceptions and cognitions of the human examiner plays a key role in forensic work. Dr. Dror has published over 100 peer-reviewed articles specifically looking at cognitive factors that medicate human expert performance, and has developed best practices in forensic work. He has provided training on 'Cognitive Factors in Forensic Decision Making' to the FBI, LAPD, NYPD, Kansas, and many other forensic laboratories across the US (as well as other countries, such as Australia, Finland, the United Kingdom and the Netherlands). Dr. Dror has received the ABP Annual Award for his accomplishments and impact in forensic science. He has also been extensively cited in National Academy of Sciences Report on Forensic Science, and was the Chair of the OSAC Forensic Human Factors group. More information is available at: www.cci-hq.com i.dror@ucl.ac.uk

References

1. Ashbaugh DR. The premise of friction ridge identification, clarity, and the identification process. J Forensic Identification 1994; 44: 499-516.
2. Forensic Regulator. Cognitive bias effects relevant to forensic science examinations. 2015. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/356003/2014.08.28_FSR_Cognitive_bias_draft.pdf.
3. Dror IE. Cognitive neuroscience in forensic science: Understanding and utilising the human element. Philos T R SOC A 2015; 370 (1674). DOI:10.1098/rstb.2014.0255
4. Nickerson RS. (1998). Confirmation Bias: A ubiquitous phenomenon in many guises. Rev Gen Psychol 1998; 2(2): 175–220.
5. Stacey RB. Report on the erroneous fingerprint individualization bombing case. J Forensic Identification 2004; 54: 706-718.
6. National Academy of Sciences. Strengthening forensic science in the United States: A path forward. Washington, DC: National Academy of Sciences, 2009.
7. Dror IE, Champod C, Langenburg G, Charlton D, Hunt H, Rosenthal R. Cognitive issues in fingerprint analysis: Inter-and intra-expert consistency and the effect of a 'target' comparison. Forensic Sci Int 2011; 208: 10-17.
8. Dror IE, Thompson WC, Meissner CA, Kornfield I, Krane D, Saks M, Risinger M. Context management toolbox: A Linear Sequential Unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. J Forensic Sci 2015; 60(4): 1111-1112.
9. Coble M. Interpretation Errors Detected in a NIST Inter-laboratory Study on DNA Mixture Interpretation in the U.S. (MIX13). Proceedings of the International Symposium on Forensic Science Error Management – Detection, Measurement and Mitigation. Washington, DC: National Institute of Standards and Technology, 2015.
10. Dror IE, Hampikian G. Subjectivity and bias in forensic DNA mixture interpretation. Sci Justice 2011; 51(4): 204-208.
11. Dror IE, Rosenthal R. Meta-analytically quantifying the reliability and biasability of forensic experts. J Forensic Sci 2008; 53(4): 900-903.
12. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. PLoS ONE 2012; 7(3): e32800.
13. Dror IE, Mnookin J. The use of technology in human expert domains: Challenges and risks arising from the use of automated fingerprint identification systems in forensics. Law, Probability and Risk 2010; 9(1): 47-67.
14. Mnookin J, Cole S, Dror IE, Fisher B, Houck M, Inman K, Kaye D, Koehler J, Langenburg G, Risinger M, Rudin N, Siegel J, Stoney D. The need for a research culture in the forensic sciences. UCLA Law Rev 2011; 58(3): 725-779.
15. Dror IE. Practical Solutions to Cognitive and Human Factor Challenges in Forensic Science. Forensic Science Policy & Management 2014; 4: 105-113.
16. Stammers S, Bunn S. Unintentional bias in forensic investigation. Houses of parliament POSTbrief PB-15. 2015. http://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PB-0015#fullreport.

Treatment of Error and Uncertainty in the Literature:

A Source of Enlightenment and Confusion

Ted Vosk

The literature produced by the forensic sciences cannot be fully evaluated without considering the purpose of forensic science itself. Its sole reason for existing is to aid in the investigation and prosecution of crimes.[1] The intended users of the work products of forensic science are, therefore, not just scientists, but, perhaps predominantly, laypeople who must be able to properly understand, apply, weigh and communicate them during the course of criminal investigations and legal proceedings. Accordingly, it is important that the literature produced by the forensic sciences not only be sound, but accessible to the law enforcement and legal professionals seeking to utilize forensic results. Moreover, it must be crafted with the expectation that it may be relied upon by lay jurors in determinations of guilt and innocence. Thus, to fully accomplish the mission for which forensic science was created, forensic literature must be sound and useful across four dimensions/contexts: scientific, investigative, legal and evidential.

The importance of peer reviewed literature is well recognized.[2-4] Given that the variety of forensic resources relied upon in the contexts set forth above, however, consideration cannot be limited to such publications. Rather, the relevant literature includes simple articles, consensus standards, government studies and books. Although some of the available resources are rigorously peer reviewed, others are subject to a lesser, quasi-peer review standard while others receive little to no peer review at all. Because the distinction is often unappreciated by lay jurors, and lawyers are advocates who often ignore it, a single unreviewed resource, such as a book intended for nonspecialists, may trump a collection of peer reviewed papers and standards to the contrary.[5] Thus, it is critical that each of these resources be considered.

Consensus standards are important because they are intended to encapsulate the state of accepted scientific practices. Nonetheless, they and government studies can pose particular risks. This arises from several factors: first, they are both stamped with the imprimatur of authority; second, the review process for standards may include consideration of factors that are not scientific in nature but relate to other, often professional or ideological, interests;[6] and third, government studies are often not rigorously reviewed and published on the basis of political/social agendas as opposed to sound science. As an example of the latter, the National Highway Traffic Safety Association (NHTSA) has published several "validation studies" of the Standardized Field Sobriety Tests (SFSTs), none of which have been peer reviewed.[7-9] The combination of these factors can lead to the acceptance of results and principles within both forensic and legal communities, as well as by fact-finders in the courtroom, that are not well supported by science. The situation is exacerbated by the fact that these materials can be admitted into evidence without any supporting expert testimony so that their content may go unchallenged.[10-12]

The contexts within which error/uncertainty are typically encountered in the forensic literature can be broken down into five categories: (1) reporting of the error/uncertainty associated with methods/results; (2) instruction on methods for determining error/uncertainty; (3) explanation of theory underlying error/uncertainty; (4) detailing of quality procedures; and (5) policy discussions. Despite the context, the importance of error/uncertainty lays in the role it plays in constraining the inferences permitted by scientific methods and results. No scientific method or result permits us to "know" the truth of a particular proposition. The best that it can do is make it more or less likely. Measures of error and uncertainty provide the means by which the conclusions supported by a result can be understood and weighed.

The literature with respect to measurement error and uncertainty is far better developed than that concerning most qualitative forensic methods with the exception of DNA analysis.[13] The available resources cover each of the above categories, target both forensic and legal communities, include both government and private forensic scientists, and are generally on par with those found throughout the general scientific community.[14-18] Their primary shortcoming is in the area of theory. The literature concerning forensic measurement evidences a bias toward methods of determining error/uncertainty with less emphasis on understanding what they actually represent, but there are exceptions.[19, 20] Although the literature with respect to qualitative methods still must improve, it has grown in recent years with some notable contributions.[21]

A common problem encountered in the forensic literature is carelessness in the use of terminology. Concepts including error, error-rate and uncertainty are often ill defined and/or misused. For example, in the area of forensic measurements the terms error and uncertainty are often used interchangeably despite the fact that they represent completely distinct metrological concepts which convey different types of information about a measurement (Metrology is the science of measurement and its application).[13, 22, 23] As a result, authors commonly discuss and describe measurement error as if it were measurement uncertainty while being unaware of the full meaning of the former and the independent existence of the latter.[24, 25] This creates confusion which slows development of sound measures of error/uncertainty and undermines expert testimony relying upon these resources as their basis of knowledge.

Another shortcoming within the forensic literature is the mischaracterization of data and conclusions within technical papers by reported error/uncertainty.[26, 27] The impacts of this are felt not just within forensic science but the courtroom where such materials are relied upon as evidence and undermine the fact-finding process. The origin of the problem is at least partly due to the fact that many subject matter experts tend to have little sophistication in experimental design and data analysis. Even when papers are submitted for peer review, the process typically relies upon reviewers with similar subject matter background and, therefore, deficits similar to those of the submitting authors. Thus, the same mischaracterizations can be found in peer reviewed as non-peer reviewed literature, albeit to a lesser extent.

A possible way of addressing this situation is to encourage multi-disciplinary authorship, including both subject matter and error/uncertainty experts on publications, and, where journals are concerned, inclusion of error/uncertainty experts as reviewers prior to publication.

An area of concern not often appreciated in the evaluation of forensic science literature is the relationship between forensic science and the law itself. If forensic science is not understood by the legal professionals who work with it, the justice system will not be able to properly incorporate or have it considered. The inaccessibility of the forensic literature to many judges and lawyers, as well as lay jurors, subverts both pretrial decision-making and courtroom fact-finding processes by rendering many legal professionals and lay jurors unable to weigh claims contained within, or based upon, the literature in a manner consistent with the science therein. Rather, they are often forced to rely upon faith in the representations of a claimed expert or the acumen of legal professionals in establishing or destroying a witness's credibility when evaluating claims within or based upon the literature, neither of which may have anything to do with the scientific itself.

On the other hand, if forensic scientists don't understand how the law relates to the goals of their activities, the methods and conclusions they publish may not be relevant to the questions of fact at issue in a legal matter. This is true even though the courts may not see the irrelevancy due to their lack of scientific sophistication. Regardless of how good the science being published is, a misunderstanding of those aspects of the law the methods and conclusions are meant to address can lead to confusion and unintended deception.[28]

Once again, multi-disciplinary authorship and review could yield far reaching gains. This time, however, the collaboration suggested is between forensic scientists and legal professionals (i.e., judges and lawyers). Although there are already excellent journals whose focus is the intersection of law and science and which regularly include both legal and forensic professionals,[29] they tend to be aimed at a more academic audience. Both forensic science and the law would benefit from inclusion of legal professionals as both frequent co-authors and/or reviewers of more practitioner oriented forensic publications.[30] Given that one of the primary consumers of forensic science work product and literature are laymen, forensic publications should not be merely technical documents written for other scientists, but explanatory in nature as well, aimed at educating those legal and law enforcement professionals who must rely upon the methods and results described.[31] Moreover, this collaboration can help to ensure that the forensic science being published is in fact addressing the questions raised within the criminal justice and investigation arenas.

Although already good in some areas and improving in most others, as a whole, the forensic literature associated with error and uncertainty has not yet attained a level where it can be considered generally reliable. Nor can it be considered accessible to all those who must rely upon it. The suggestions offered herein address some of the underlying causes but not all. Use of non-peer reviewed publications as a means of self-marketing and special interest bias in the drafting of consensus standards and government forensic publications remain significant issues that must be addressed.

Ted Vosk, J.D., is a criminal defense attorney and legal/forensic consultant. Homeless as a teenager, Mr. Vosk became a Goldwater Scholar, graduating with honors in Theoretical Physics and Mathematics from Eastern Michigan University, before studying physics in the Ph.D. program at Cornell University and obtaining his J.D. from Harvard Law School. Since then, he has been a prosecutor, public defender, the Acting Managing Director of an NSF Science and

Technology Center and Vice President of Celestial North which was recognized by Astronomy Magazine for Excellence in Astronomy Outreach. He has been part of some the most significant forensic science cases in Washington State over the past 15 years and led the fight resulting in significant reforms in the practices of the State's Toxicology Lab, for which he received the President's Award from the Washington Association of Criminal Defense Lawyers and the only Certificate of Distinction ever awarded by the Washington Foundation for Criminal Justice. A Fellow of the American Academy of Forensic Sciences and member of Mensa, he is an international speaker and consultant. He is a peer reviewed author whose work has been published in both legal and scientific media, and is co-author of the books *Forensic Metrology: Scientific Measurement and Inference for Lawyers, Judges, and Criminalists* and *Defending DUIs in Washington*. tvosk@comcast.net

References

1. Vosk T, Emery A. Forensic Metrology: Scientific Measurement and Inference for Lawyers, Judges, and Criminalists. London: CRC Press, 2014.
2. Alberts B, Hanson B, Kelner KL. 2008. Reviewing Peer Review. Science 2008; 321: 15.
3. National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press, 2009.
4. *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).
5. *State v. King County Dist. Court West Div.*, 307 P.3d 765 (2013).
6. National Safety Council's Alcohol, Drugs and Impairment Division. Executive Board Minutes. October 19, 2014. Grand Rapids, MI.
7. Burns M, Anderson E. A Colorado Validation Study of the Standardized Field Sobriety Test (SFST) Battery. Washington, DC: National Highway Traffic Safety Administration, U.S. Department of Transportation, 1995.
8. Burns M, Dioquino T. A Florida Validation Study of the Standardized Field Sobriety Test (SFST) Battery. Washington, DC: National Highway Traffic Safety Administration, U.S. Department of Transportation, 1997.
9. Stuster J, Burns M. Validation of the Standardized Field Sobriety Test Battery at BACs Below 0.10 Percent. Washington, DC: National Highway Traffic Safety Administration, U.S. Department of Transportation, 1998.
10. Federal Rule of Evidence 803(8), (17) & (18).
11. *Nordstrom v. White Metal Rolling and Stamping Corp*., 453 P.2d 619 (1969).
12. Philo H. Use of Safety Standards, Codes and Practices in Tort Litigation. Notre Dame Law Rev 1965; 41(1): 1-12.
13. The Joint Committee for Guides in Metrology. International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM). Geneva, Switzerland: International Organization for Standardization, 2008.
14. Vosk T. Measurement Uncertainty: Requirement for Admission of Forensic Science Evidence. In: Wiley Encyclopedia of Forensic Science. Hoboken: John Wiley and Sons, Inc., 2015.
15. Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG). Measurement Uncertainty for Weight Determinations in Seized Drug Analysis. SWGDRUG, 2011.
16. Sklerov J, Couper F. Calculation and verification of blood ethanol measurement uncertainty for headspace gas chromatography. J Anal Toxicol 2011; 35(7): 402-410.

17. Brach R, Dunn P. Uncertainty Analysis for Forensic Science. Tucson, AZ: Lawyers & Judges Publishing Company, Inc., 2009.
18. Jones W. Dealing with Uncertainty in Chemical Measurements. Newsletter of the International Association of Chemical Testing 2003; 14(1): 6-11.
19. Vosk T. Measurement Uncertainty. In: The Encyclopedia of Forensic Sciences. Waltham: Academic Press, 2013.
20. Wallace J. Ten Methods for Calculating the Uncertainty of Measurement. Science & Justice 2010; 50(4): 182-186.
21. Neumann C, Evett IW, Skerrett J. Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm. J R Stat Soc Ser A-G 2012; 175(2): 371-415.
22. The Joint Committee for Guides in Metrology. Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement (GUM). Geneva, Switzerland: International Organization for Standardization, 2008.
23. Kacker R, Sommer KD, Siebert B. Evolution of modern approaches to express uncertainty in measurement. Metrologia 2007; 44: 513-529.
24. Vosk T. Uncertain Justice: Measurement Uncertainty and the Discovery of Truth in the Courtroom. The Judge's Journal 2015; 54(3): 8-11, 39.
25. ANSI-ASQ National Accreditation Board. Guidance on Uncertainty and Traceability for Forensic Agencies. ANSI-ASQ National Accreditation Board, 2015.
26. Vosk T, Bianchi G. Field Sobriety Testing and Driver Impairment: Linked or Not? In Understanding DUI Scientific Evidence. United States: Thomson Reuters/Aspatore, 2009; 123-185.
27. Rubenzer S. The Standardized Field Sobriety Tests: A Review of Scientific and Legal Issues. Law Human Behav 2008; 32(4): 293-313.
28. Vosk T, Forrest ARW, Emery A, McLane L. 2014. The measurand problem in breath alcohol testing. J. Forensic Sci 2014; 59(3): 811-815.
29. Journal of Law, Probability & Risk. http://lpr.oxfordjournals.org/.
30. Hlastala M, Polissar N, Oberman S. Statistical evaluation of standardized field sobriety tests. J Forensic Sci 2005; 50(3): 662-9.
31. Kane G. The methodological quality of three foundational law enforcement drug influence evaluation validation studies. J Negative Results in Biomedicine 2013; 12: 16-28.

Impact of Forensic Literature on the Admissibility Process

Michael T. Ambrosino

Washington, D.C. is the only jurisdiction where the U.S. Attorney's Office prosecutes federal crimes in the United States District Court and local crimes in the Superior Court for the District of Columbia. Over the past few decades, admissibility of forensic sciences has been governed by the *Daubert* standard in District Court and the *Frye* standard in Superior Court. Though these standards profess to employ differing admissibility standards, the practical application has varied little in this jurisdiction.

When confronted with a challenge to the admissibility of a forensic science, a prosecutor must ask: what is the state of the forensic science? Is this a field or technique that is settled, evolving, or completely new? The answer lies with the *relevant scientific community*. But who comprises this scientific community? Who is relevant and who is irrelevant? In deciding whether to admit forensic evidence over the objection of the defense, judges have repeatedly looked to the following groups for guidance:

- Practitioners in the field;
- Persons who publish *peer-reviewed* scientific literature;
- Persons who perform developmental and laboratory validation; and
- Members of national and international scientific associations that set guidelines and standards.

The degree of peer-reviewed literature and validation varies significantly among forensic disciplines. In 2009, the National Academy of Sciences (NAS) report was extremely critical of the pattern matching disciplines. This sparked a series of challenges by the defense bar in Washington, D.C. to the admissibility of latent fingerprint identification and firearms and toolmark identification. At the time, there was very little validation that had been conducted in the field of latent fingerprint identification. Despite a dearth in validation, challenges to the admissibility of latent fingerprint identification were repeatedly rejected by the judges, largely due to the scientific studies research -- such as the twin studies -- that supported the theory that fingerprints are unique. Although the NAS report had little impact on the admissibility of latent fingerprint evidence, it did spark the relevant scientific community to conduct validation and to explore the application of statistics to the field.

Unlike latent fingerprint identification, there was no biological underpinning to the theory that certain firearms created individualized markings on casings and bullets. However, firearms and toolmark identification did not suffer from a dearth of validation studies at the time the NAS report was issued. Numerous studies demonstrated that trained firearms examiners could identify individual characteristics in bullets and casings fired from sequentially manufactured firearms. The judges, and eventually the D.C. Court of Appeals, rejected efforts by the defense bar to exclude firearms and toolmark identification evidence.

Despite the continued admission of firearms and fingerprint evidence, the U.S. Attorney's Office did limit the certainty with which its experts expressed a match. For example, experts no longer expressed identifications with "absolute," or "100%" certainty. Nor did they make identifications to the exclusion of all other firearms/fingerprints in the world. Rather, experts would identify a casing/bullet to a firearm or a latent print to a known exemplar "to a reasonable degree of scientific certainty." Moreover, if these fields ever develop statistical models to express the certainty of a match, those will be incorporated into the presentation of evidence.

Ironically, when the defense bar initially objected to the admissibility of PCR-STR DNA testing in the late 1990s, it asserted that DNA was "junk science" and did not enjoy general acceptance within that particular community of scientists as was the case with more established forensic disciplines such as Latent Fingerprint Identification and Firearms and Toolmark Identification. Not surprisingly, these arguments failed. Although the use of forensic DNA as an investigative tool did not become commonplace until the 1990s, it has attained status as the "gold standard" due to the scientific literature and validation that underlies each new advancement. Judges have had little patience with challenges to the admission of subsequently developed methodologies within the field of DNA – such as Mitochondrial, Y-STRs, and MiniSTRs – given that these methodologies are also grounded in validation and widely accepted in the peer-reviewed published literature. Peer-reviewed scientific literature was also critical in evaluating and admitting various statistical calculations to express the significance that a DNA profile developed from an item of evidence matched the profile of a known person.

Recently, there has been a concerted effort by some members of the defense bar to have the DNA results manipulated in a way that is not in accordance with accepted scientific standards. In *U.S. v. Cardell Torney*, an unknown male profile was developed from the anal swab of a victim who was abducted by a stranger at her front door and raped in a nearby yard. The unknown male profile was uploaded into the National DNA Database (NDIS) and matched against a convicted offender. The prosecution verified the DNA match to the defendant. The defense subsequently hired NMS labs to conduct DNA testing for the defense. NMS put too much DNA into the amplification process. The saturated amplification resulted in data that was beyond the linear range of the genetic analyzer. The data was so overblown that it created artifacts at low levels that the NMS expert attempted to pass off as unknown contributors. The practices of NMS were so out of line with accepted scientific practices that the President of Promega Corporation sent a representative to testify that NMS had misused Promega's STR kit PowerPlex 16 HS. Moreover, world-renown DNA expert Dr. Bruce Budowle demonstrated to the court how NMS's practices were not supported by its own validation of the Promega kit.

In *U.S. v. Demonata,* the defense provided the prosecution with notice of defense DNA testing results on the eve of trial. Examination of the defense DNA data revealed that the defense expert had conducted Low Copy Number (LCN) DNA testing that had not undergone proper validation. The government objected to the admission of the evidence as scientifically unsound and requested an emergency *Frye* admissibility hearing. The defense asserted that it was simply a battle of the experts that should play out before the jury. The court disagreed and scheduled an emergency *Frye* hearing. The defense subsequently withdrew its effort to utilize LCN testing results a day prior to the admissibility hearing.

In *U.S. v. James Dorsey,* the defense opted to conduct DNA testing at DNA Diagnostics Center (DDC). The defense instructed DDC to conduct STR and Y-STR testing. However, the defense only disclosed the STR testing results to the prosecution. During the prosecution's cross-examination of the DDC DNA analyst, the Y-STR testing came to light. The analyst, however, asserted that she could not recall whether the Y-STR profile developed from the victim's jacket was consistent with defendant's Y-STR profile. According to the DNA analyst, the defense attorney had instructed her not to disclose the results of the Y-STR testing. The prosecution subsequently moved the court to compel production of the Y-STR testing results from the defense. When defense counsel asserted that he did not possess the results, the court signed a subpoena for DDC to produce the Y-STR testing results. The defense subsequently moved to withdraw the testimony of its own DNA expert which made production of the Y-STR results moot.

In the fall of 2014, Dr. Bruce Budowle was hired by the U.S. Attorney's Office to conduct likelihood ratios in *U.S. v. Tavon* Barber. Upon review of the data, Dr. Budowle identified an issue regarding the interpretation of data involving mixture profiles by DNA analysts at the Department of Forensic Sciences (DFS). This prompted a further review of other pending prosecutions in which DFS analysts had interpreted DNA mixtures. After Dr. Budowle identified a number of similar issues in other prosecutions, the U.S. Attorney's Office convened a panel of experts to conduct a comprehensive review of all past convictions and pending cases where DFS analysts had interpreted DNA mixtures. Prior to the completion of the review, the U.S. Attorney's Office pulled all DNA casework from DFS and hired Dr. Budowle and Dr. Frederick Bieber to conduct an audit of the DNA mixture interpretation by DFS analysts. The Mayor of Washington, D.C. subsequently hired ANSI-ASQ National Accreditation Board (ANAB) to audit the practices of the DFS's DNA Unit. In April 2015, Drs. Budowle and Bieber issued a report that was highly critical of DNA mixture interpretation at DFS. Accepted scientific literature weighed heavily in the evaluation of practices at the DFS. For example, the Budowle-Bieber audit contained numerous citations to John Butler's most recent text book on DNA mixture interpretation. A day after the issuance of the Budowle-Bieber audit, the ANAB auditors issued a report ordering the DNA Unit to be shut down due to a series of non-compliances with accepted scientific practices.

These examples demonstrate that even the gold standard of forensics is not immune to attack when practitioners stray from accepted scientific practices.

Michael T. Ambrosino, JD, has served for the past five years as the Special Counsel for DNA and Forensics at the U.S. Attorney's Office for the District of Columbia. He is responsible for overseeing the use of forensic evidence by DC prosecutors and all forensic functions at the U.S. Attorney's Office, including briefing and arguing all forensic admissibility hearings of first impression. Prior to serving as Special Counsel to the U.S. Attorney, Mr. Ambrosino spent over a decade handling homicide cases in which he handled numerous forensic matters of first impression, including the admissibility of STRs and the first DNA cold hit case in the District of Columbia. Over the past two decades, Mr. Ambrosino has argued over forty forensic admissibility hearings. michael.ambrosino@usdoj.gov

Section 2: Interpretation and Assessment

Discussion Summary

The panelists John Lentini, Michael Shermer, Orla Smith, Itiel Dror, Ted Vosk, and Michael T. Ambrosino, discussed the necessity of validation data and error analysis, whether other areas have the same widespread issues as fire investigation, and the relationship of peer reviewed journals with new alternative publishing services. Ambrosino talked about the automated methods that can now be used for analysis of DNA mixtures and degraded DNA; his conclusion was that he could not use them in court unless the companies are transparent about their validation data. It was agreed that some methods aren't amenable to error analysis, that analyses such as shoeprint analysis may have to be considered "more art than science," and that some forensic results are still unable to be quantified. However, Mark Stolorow mentioned that NIST had just approved a new Center of Excellence, established to address issues such as probabilistic determinations and models for such calculations, which may lead to new advances in this area. Lentini discussed how rampant "mythology" is unique to fire analysis; however, lack of scientific training in practitioners can also be seen in other areas such as ballistics and fingerprints. He mentioned that fire investigation is at risk of being discounted by the forensic community, highlighting one of the negative effects that bad literature can have. Finally, Smith talked about how most publishing houses have rigorously peer reviewed open access journals, but it is yet to be seen how alternative review models will affect the status of those houses.

Section 3: Policy Implications

Plenary III.
Policy Implications of Inadequate Literature

Ronald N. Kostoff

The following is an outline on policy implications of inadequate literature. An expanded version is available online.[1] The characteristics are identified as follows: A) Those characteristics of literature that are adequate to support policy; B) How each of these characteristics can become degraded to an inadequate literature; C) Identification of incentives for inadequate literature; and D) Policy implications of inadequate literature.

Literature characteristics adequate to support policy

1) All critical research problems necessary for credible policy addressed/funded.
2) All research performed is credible and high quality.
3) All research findings submitted for publication.
4) All high quality research submissions published.
5) All published articles available to general public.
6) All accessible articles easily retrieved.

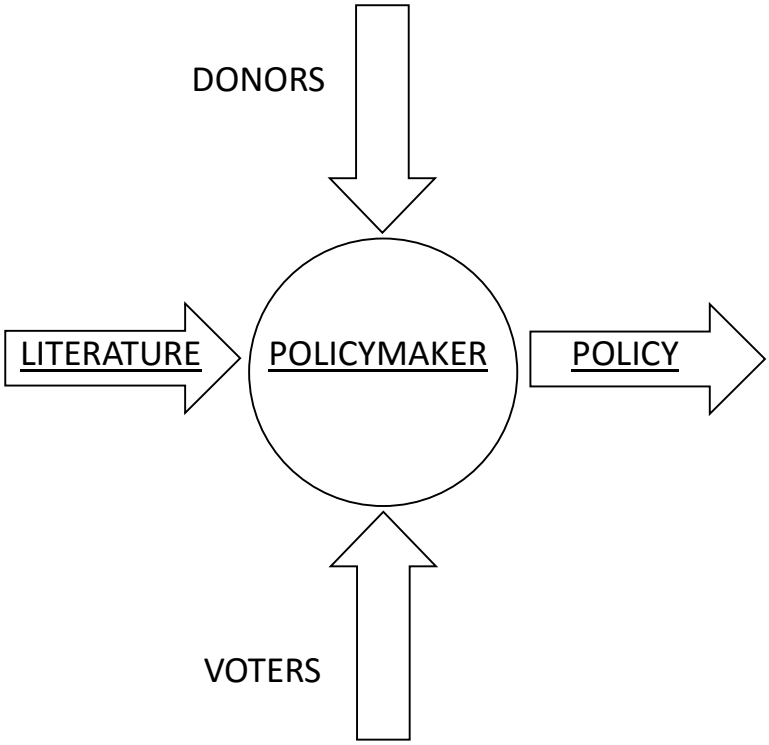
Inadequate literature

1) Some critical research problems not addressed/funded (potentially because of insufficient funds, poor research topic selection, industry pressure [commercial sensitivity], and/or government pressure [political sensitivity]).
2) Research findings not submitted for publication (potentially because of classification for national security or other purposes, deliberate suppression of findings [Negative findings, Adverse events, Unethical research, Commercial sensitivity, Political sensitivity], organizational proprietary, and/or no organizational or individual publishing tradition).
3) Low quality research published (potentially because of poor peer review, and/or contributes to Editor's pre-determined agenda) and high quality research not published (potentially because of poor/biased peer review, non-contribution to increasing journal Impact Factor, and or non-contribution to Editor's pre-determined agenda) and manufactured research published (documented in Merchants of Doubt; Science for Sale[2,3]; purpose is to counter publications showing adverse effects and to sow confusion, not allowing consensus required for policy).
4) Published research not easily accessible (potentially because of obscure media, high paywalls, and/or poor search algorithms).

Incentives for inadequate literature

1) Industry (conceal adverse effects of products and services).

2) Government (at upper levels, support corporate or individual large donor interests; at lower levels, support upper level interests for advancement; lay groundwork for future industry employment).
3) Journal Editors (many journals sponsored by professional societies or supported by advertising; professional societies supported strongly by industry; sponsors/advertisers not interested in articles adverse to industrial products[4]).
4) Research Performers (grants and publications important for advancement; work in areas for which funds are available; important to satisfy sponsors to maintain grants; important to satisfy Editors to get published; and/or lay groundwork for future industry employment and/or grants).



The above figure shows the main drivers of policy. The published literature, technical or otherwise, is only one of the drivers, and as will be shown, may not be the most important driver in many cases. The policymaker is assumed to be a government employee, whose organization is headed by a political appointee at the highest level, and who is therefore influenced by major political determinants: campaign donors and the electorate.

Policy Implications of inadequate literature
Option 1) The topical area is **non-sensitive commercially or politically** (e.g., weather satellite research; age of universe research). Donors and voters would agree with policy dictated by adequate literature; donors and voters agree with policy dictated by inadequate literature; policy reflects literature. An **inadequate literature results in inadequate policy**.

Option 2) The topical area is **sensitive commercially or politically** (e.g., climate change amelioration; EMF health impacts[5]). In this case, **donors and voters would *disagree* with policy dictated by adequate literature**. The donors are driven by profit, and the voters are

addicted to technology. **Thus, donors and voters agree with policy dictated by *inadequate*** literature. For e.g. EMF health impacts, the policy that a hard reading of the credible technical literature might require (severe restrictions on the use of wireless communications, etc.) would not be acceptable to the vast majority of donors and voters. Thus, the policy reflects the interests of the donors and voters, not the dictates of an adequate technical literature. In other words, the policy reflects that dictated by an inadequate literature.

Option 3) The topical area is **sensitive commercially and/or politically** (e.g., exposures/treatments that cause disease[6]). In this case, **donors would *disagree* with policy dictated by adequate literature**, whereas the **voters would *agree* with policy dictated by adequate literature**. The donors are driven by profit, whereas the voters are driven by the benefits of technology. Unlike the previous case, the voters are not addicted to the technology, since its application may be unpleasant in many cases. The donors still agree with policy dictated by inadequate literature, whereas the **voters agree with policy dictated by inadequate literature, *only because they believe it is adequate***. This means that some literatures may be highly manufactured to maintain voter support. The policy reflects donors, not adequate technical literature. In other words, the policy reflects inadequate technical literature.

Conclusion

The published technical literature is inadequate for myriad of reasons, and the degree of inadequacy is unknown, and may be unknowable. The fraction of inadequacy due to deliberate misinformation is unknown, and may be large.

Ronald N. Kostoff, Ph.D., is a Research Affiliate at the School of Public Policy, Georgia Institute of Technology. Dr. Kostoff received his Ph.D. in Aerospace and Mechanical Sciences from Princeton University in 1967. He has worked for Bell Laboratories, the Department of Energy, the Office of Naval Research, and MITRE Corp. He has authored over 200 articles, served as Guest Editor of four journal Special Issues, obtained two text mining system patents, and presently is a Research Affiliate at Georgia Institute of Technology. Technical areas published include Aerospace (aerothermodynamics, high speed gas dynamics, rarefied gas dynamics, orbital mechanics); Energy (fusion, fission, energy system economics); Technical Assessment (research evaluation; peer review; S&T roadmaps, country S&T [China, India, Finland, Mexico, Brazil]); Information Technology (text mining [nanotechnology, mass spectrometry, fractals, nonlinear dynamics, electrochemical power sources, all power sources, aircraft, abrupt wing stall, high-speed compressible flow, military-relevant S&T, anthrax, EMF health impacts, neuropsychology, prevention and reversal of chronic disease], literature-based discovery [Multiple Sclerosis, Parkinson's Disease, Raynaud's Phenomenon, Cataracts, SARS, Vitreous Restoration, Chronic Kidney Disease, Water Purification]).
ronald.kostoff@pubpolicy.gatech.edu

References

1. Kostoff RN. Pervasive Causes of Disease. Georgia Institute of Technology 2015. http://hdl.handle.net/1853/53714.
2. Oreskes N, Conway EM. Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming. London: Bloomsbury Press, 2011.

3. Lewis DL. Science for Sale: How the US Government Uses Powerful Corporations and Leading Universities to Support Government Policies, Silence Top Scientists, Jeopardize Our Health, and Protect Corporate Profits. New York: Skyhorse Publishing, 2014.
4. Slesin L. "Radiation Research" and the cult of negative results. Microwave News. 2006 July. http://microwavenews.com/RR.html.
5. Kostoff RN, Lau CGY. Combined biological and health effects of electromagnetic fields and other agents in the published literature. Technol Forecast Soc 2013; 80(7): 1331-1349.
6. Kostoff RN, Patel U. Literature-related discovery and innovation: Chronic kidney disease. Technol Forecast Soc 2014. doi: mhttp://dx.doi.org/10.1016/j.techfore.2014.09.013.

A Quality and Gap Analysis: An AAAS Forensic Science Literature Project

Deborah Runkle

Preface

This paper describes a project that is being conducted by the American Association for the Advancement of Science (AAAS). The project is designed to respond to an important societal need – improving the scientific underpinnings of forensic practices that are introduced at criminal trials and often mean the difference between guilt and innocence, even life or death.

Background

Valid and reliable forensic tools are essential to the criminal justice system to help determine guilt or innocence. Nevertheless, there have long been assertions that the forensic sciences, other than DNA analysis, are neither valid nor reliable. In fact, in some cases, reports and testimony based on substandard science have contributed to the convictions of individuals later proved innocent through DNA testing. The Innocence Project has successfully exonerated over 300 individuals who were wrongly convicted, with faulty forensic evidence playing a part in 50% of the cases. (In many instances, more than one factor contributed to the wrongful conviction; erroneous eyewitness identification is the most common factor.)[1] Brandon Garrett, law professor at the University of Virginia, claims that criminal justice relying on shaky science has created a "mass disaster."[2] Professor Garrett made these remarks following the FBI's disclosure that it is reviewing 2,500 cases where hair-sample matches were presented at trial.[3] Preliminary results of that review show that between 1985 and 1999, of 268 cases nationwide, 96% resulted in erroneous testimony, and 32 of the defendants received death sentences, some of whom were executed.[4]

Criminal defendants are disadvantaged in court because, typically, the forensic scientists analyzing the evidence are part of law enforcement.[5] Criminal defendants are often indigent and cannot afford a lawyer, let alone their own expert. The public defenders or contract lawyers representing these individuals often lack the necessary knowledge and skills to challenge the state's expert witnesses; further, in most state jurisdictions, they must get approval from the presiding judge in order to retain an expert.[6] Also, because nearly all forensic fields were developed in the context of crime solving, not academic science, independent research is virtually non-existent.

The Supreme Court's 1993 *Daubert* decision initially seemed to offer a remedy to the disadvantaged defendant.[7] The Court ruled that under Rule 702 of the Federal Rules of Evidence a "trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant but reliable." In short, federal judges have an obligation to be "gatekeepers," separating good science from dubious science in the pre-trial stage. Although some states still rely on the *Frye* standard, most states have adopted *Daubert* or something close to it.

Attorneys representing defendants in civil cases have taken advantage of the door the Court opened for them, challenging the admissibility of plaintiffs' evidence for failing to meet the standards set forth in *Daubert*. However, unlike their counterparts in civil practice, attorneys representing criminal defendants make relatively few challenges to the state's evidence in criminal trials.[8] This finding was confirmed by an important 2009 report from the National Academy of Sciences/National Research Council (NAS), which concluded that in criminal cases "forensic science evidence is not routinely scrutinized pursuant to the standard of reliability enunciated in *Daubert*."[7] If the defense bar fails to challenge prosecutors' evidence, it follows that "*Daubert* did not have the same effect in criminal litigation [as in civil litigation]."[9] Even where the defense mounts a challenge, the disputed forensic evidence is more often than not admitted.[10]

Why did this important ruling by the Supreme Court fail to make an impact on criminal justice? There are several plausible answers. First, as noted above, attorneys representing indigent defendants may not be as sophisticated or have the resources that are available to attorneys in civil practice. Second, judges may be reluctant to exclude the state's forensic evidence because of the disruption in law enforcement that might ensue. Third, judges may be influenced more by precedence regarding the admission of forensic evidence than by the standards set forth in *Daubert*. Further, in those relatively few cases where prosecution evidence was challenged under the *Daubert* standards, judges may have over-relied on the Court's instruction that these standards were meant to be "flexible."

On the assumption that "the judicial system has failed" to provide criminal defendants with an opportunity to probe the prosecution's forensic evidence in order to evaluate its reliability, Peter Neufeld, founder and Co-Director of the Innocence Project, says that "other remedies must be found further 'upstream' …"[6] Congress agreed with this assessment and, based on the Science, State, Justice, Commerce, and Related Agencies Appropriations Act of 2006,[6] authorized the NAS to conduct a study on forensic science.

In 2006, the NAS formed a committee to examine several aspects of forensic science, and in 2009, issued its report, "Strengthening Forensic Science in the United States: A Path Forward." Importantly, the committee agreed with criticisms of the judicial system regarding criminal defendants, stating that "Review of reported judicial opinions reveals that…forensic science evidence is not routinely scrutinized pursuant to the standard of reliability enunciated by *Daubert"* and "*Daubert* has done little to improve the use of forensic evidence in criminal cases."[11]

The report's most dramatic conclusion was that some forensic science as currently practiced has "little systematic research to validate the discipline's basic premises and techniques." The report further noted that because most forensic analysis was developed in crime labs, it has "never been exposed to stringent scientific testing." This state of affairs is largely due to the fact that "[t]he broader scientific community generally is not engaged in conducting research relevant to advancing the forensic science disciplines." For that reason, the first recommendation of the NAS committee included an effort to "promot[e] scholarly, competitive peer-reviewed research and technical development…."[8]

The NAS report was a good start on improving forensic science and, coming from such a prestigious organization, has been widely quoted and influential. Nevertheless, in its overall critique of forensic science, it did not specifically point to the areas where the practices are supported by sound research and where they are not. And while it called for research from the "scientific community," it did not provide a research agenda to aid that community, including funding agencies.

In an effort to follow through on the NAS report, the National Science and Technology Council (NSTC)[8] Committee on Science established a Subcommittee on Forensic Science (SoFS) in July 2009. SoFS' charter defines its purpose as "advis[ing] and assist[ing] the…Executive Office of the President on policies, procedures and plans related to forensic science in…criminal justice…." SoFS members were told to emphasize "developing practical and timely approaches to enhancing the validity and reliability…in forensic science…." In May 2012, the SoFS charter was reauthorized, with an "encouragement" to "create a prioritized forensic science research agenda through a comprehensive gap analysis or other appropriate means…."[12]

In fulfilling its responsibilities, SoFS appointed five Interagency Working Groups (IWGs). Among the IWGs, one – the Research, Development, Testing, and Evaluation IWG (RDTE IWG)—was tasked with addressing that "encouragement." RDTE IWG members realized they could create neither a gap analysis nor a research agenda without first ascertaining what literature already existed. Members drew up a set of questions for ten forensic fields: (1) bite marks, (2) bloodstain patterns, (3) digital evidence, (4) fiber evidence, (5) fire investigation, (6) firearms and tool marks, (7) footwear and tire tracks, (8) hair evidence, (9) latent fingerprints, (10) paints and coatings. The answers to these questions were deemed critical in determining whether or not a particular discipline was valid and reliable. The members sent the questions to relevant forensic entities representing these fields, asking them to identify books, articles or other sources that supported the questions. The result of this effort, while not a gap analysis, was an important step on the way to that goal, providing future investigators with an annotated bibliography for each of the ten fields. In May 2014, the NSTC issued a report – "Strengthening Forensic Science" – based on the conclusions reached by each of the five IWGs that together constituted the SoFS. The report stated that the RDTE IWG had "pursued the identification of foundational research that can be mapped to specific principles across the various disciplines of forensic science."[13]

Coincident with the SoFS activities, Congress took note of the problems with current forensic practice. Specifically, Senator Jay Rockefeller, Chair of the US Senate Committee on Commerce, Science, and Transportation, held three hearings on forensics. The first in the series was held in December 2011,[14] the second in May 2012,[5] and the last hearing in June 2013.[15] At the lead-off hearing, Senator Rockefeller noted that most Americans obtain their understanding of forensic science from popular TV shows like *CSI* and think that forensic science is "nearly infallible, always conclusive." He went on to say "the reality is far from this depiction." At the follow-up hearing in 2012, he said he found it disturbing that a witness at the earlier hearing had testified that many forensic fields lack a "culture of science." Senator Rockefeller later introduced the Forensic Science and Standards Act of 2012. While the Act did not become law, it

recommended that the federal government establish a "national research program to improve, expand, and coordinate Federal research in the forensic sciences."[16]

Although there was no companion bill from the House and Senator Rockefeller's bill never became law, the federal government further signaled its interest in forensic science by the establishment of the National Commission on Forensic Science by the Department of Justice and the National Institute of Standards and Technology (NIST) in April 2013. The Commission's responsibilities included the "identif[ication] and assess[ment of] the current and future needs of the forensic sciences to strengthen their disciplines…"[17] At the Commission's first meeting, in February 2014, John Holdren, Assistant to the President for Science and Technology and Director of the White House Office of Science and Technology Policy, noted that the Commission's recommendations will "help ensure that the forensic sciences are supported by the most rigorous standards available – a foundational requirement in a nation built on the credo of 'justice for all.'"[18] And Under Secretary of Commerce for Standards and Technology Patrick Gallagher said the Commission's purpose is to "help ensure that forensic science is supported by the strongest possible science-based evidence gathering, analysis and measurement."

Despite the high hopes for the Commission, at its first public meeting in February 2014, several members commented that the appointment of the Commission might be premature. That is, it is pointless to recommend requirements for accreditation, training, and certification given the uncertain status of the sciences themselves. Why train forensic scientists to competently perform unreliable and invalid tests? These Commissioners argued that an analysis of the underlying scientific and technical foundations for the forensic sciences should have precedence.

AAAS Project

For more than a decade, serious concerns have been raised about the scientific basis for the forensic fields frequently relied on in criminal trials, yet there was no systematic evaluation of the foundational literature underlying forensic practice. Within months of that first Commission meeting, AAAS began an effort, funded by the Laura and John Arnold Foundation, responsive to the concerns raised regarding the state of the science.

This project will evaluate the quality of the studies the forensic community relies on to support their practices and, where the scientific underpinning of these practices falls short, to recommend areas requiring further study. This "gap analysis" will produce a research agenda to serve as the basis for arriving at forensic methods that will inspire greater confidence in our criminal justice system. Although this analysis has been long needed, and many in the public and private sphere have urged its implementation, the AAAS project will be the first attempt to fill this void. At long last, there will be a rigorous evaluation of current forensic tools and methods, enabling justice to proceed as it should. Importantly, there will be a research agenda to guide scientists and funders.

AAAS will analyze the literature for the 10 areas chosen by the SoFS IWG, set forth in footnote 24 above. To guide this effort, AAAS staff has appointed a distinguished Project Advisory Committee to advise them on all aspects of the project: Martha Bashford, Chief, S0065 Crimes Unit, New York County District Attorney; Shari Diamond, Professor of Law and Psychology, Northwestern University School of Law; Itiel Dror, University College of London;

Jules Epstein, Professor of Law, Widener University School of Law; Barbara Hervey, Judge, Texas Court of Criminal Appeals; Gilbert Omenn, Director, Center for Computational Medicine and Informatics, University of Michigan; Jeff Salyards, Director, Defense Forensic Science Center; and Hal Stern, Professor of Statistics, University of California, Irvine. Working Groups are being established for each of the 10 areas, with each Working Group consisting of up to 6 individuals. We anticipate that some forensic practitioners may "push back" on any critiques of their work, so one forensic practitioner in the field under study will be included in each Working Group. The other Working Group members will be basic scientists or other technical experts appropriate to each forensic area, for example, physicists or statisticians.

Early in its deliberations, the National Commission on Forensic Science heard testimony from several speakers about possible cognitive factors that can create "bias" in forensic analyses and the need to address them. For example, one speaker, Deborah Boehm-Davis, Dean of the College of Humanities and Social Sciences at George Mason University, testified that "selective focus helps the brain interpret information but it can also prevent information from being recognized. Therefore, bias is a normal part of brain function and the process by which humans gather and interpret information."[19] To account for this phenomenon, cognitive scientists and academics with expertise in human factors will be appointed to each Working Group.

Each Working Group will be sent the annotated bibliographies developed by the IWG described above. Because these bibliographies are several years old, staff and members of the Working Groups have updated them. The bibliographies consist of the questions identified by the IWG, with the articles/books/reports that practitioners selected as responsive to the questions. The bibliographies have up to 25 questions apiece and often more than 100 articles. Recognizing that no reviewer can evaluate each article, staff has directed the Working Group members to select for evaluation those that are compatible with their expertise. Each member will complete an evaluation template drafted by staff, asking about the strength or weakness of the article reviewed and whether it "answers" the question to which it relates.

Although AAAS will evaluate all ten of the forensic fields considered by the RDTE IWG, with the advice of the Advisory Committee three fields are being investigated in the first phase: fire investigation, firearms and tool marks, and latent fingerprints. Below are some of the questions associated with each of the fields:

Fire investigation

- What literature exists that describes how basic science in the physics and chemistry of fires is translated into the practice of burn pattern analysis for practitioners?
- What is the literature that describes the key investigative issues that must be considered when performing burn pattern analysis and arson investigation at the crime scene?
- What literature exists that describes studies on understanding how cognitive bias may affect burn pattern analysis and arson investigations?

Firearms and tool marks

- What literature documents the scientific domains used to inform the foundations of firearm/tool mark analysis?

- What research has been completed, if any, to determine the threshold for identification (individualization)?
- What studies have been performed to determine the error rates in firearm and tool mark analyses?

Latent fingerprints

- What scientific literature describes how distinct or similar fingerprints are across the overall population, related individuals or identical twins?
- What scientific literature establishes the key sources of bias and characterizes the effectiveness or measures to mitigate the bias?
- What scientific literature characterizes the effect of analyst qualifications/experience on fingerprint matching accuracy?

A report will be issued for each of the disciplines studied. The reports will identify which studies currently in the literature provide solid support for the forensic field and which do not, thereby producing a list of issues that require further research and who – physicists, engineers, biologists – should conduct that research. There will be two versions of each report, a technical version and a version in layman's language. The latter version will help enable the broader public to understand the important issues at stake in this project and the conclusions reached by the Working Groups.

Conclusion

The recommendations emerging from this project promise to be transformative to the criminal justice system. Given the demand for this type of project from both public and private domains, along with the stature AAAS has in the scientific community and beyond, the reports coming forth from the project are likely to attract wide attention. AAAS views these reports as a way to encourage basic, traditional scientists to pursue the research topics raised in the reports and funding agencies to support these scientists. The reports will be a contribution to injecting a "culture of science" into forensic research. The results of this project, scheduled to be completed in 2017, are expected to transform forensic science research and to improve the practice of forensic science, while boosting the public's confidence in the U.S. criminal justice system.

Deborah Runkle is a Senior Program Associate at the American Association for the Advancement of Science (AAAS), Associate Staff Director for the National Conference of Lawyers and Scientists, a joint standing committee of AAAS and the Science & Technology Law Section of the American Bar Association (ABA) and is a member of the Leadership Group of that Section. At AAAS Ms. Runkle is also the Staff Liaison for the AAAS Scientific Freedom and Responsibility Award, and provides staff support to the AAAS Committee on Scientific Freedom and Responsibility. She is also on the Board of Trustees for the Association for the Assessment and Accreditation of Animal Laboratory Care, Int. Her policy interests include issues at the intersection of science and law and science and society. She is currently the manager for a project, funded by the Laura and John Arnold Foundation, evaluating the literature in 10 forensic disciplines. She is also managing projects related to online patient registries (funded by the Agency for Healthcare Research and Quality), legal issues with mobile health technologies (funded by the Robert Wood Johnson Foundation), and a series of judicial educational seminars on "Emerging Issues in Neuroscience" (funded by the Dana Foundation).

She was the co-organizer of the ABA-AAAS webinar series on Neuroscience and the Law, a member of the planning committee of the National Institute of Canada's "Neuroscience in the Courtroom" program, and a member of the National Institute of Justice's Focus Group for future research directions for forensic sciences. She was a co-organizer for the AAAS-Food and Drug Law Institute's Colloquia on Personalized Medicine, co-edited *Personalized Medicine: Prescriptions and Prospects*, and was a member of the planning committee for two national conferences on patient safety. She has organized symposia and seminars at the annual meetings of the AAAS, ABA, Society for Neuroscience, and the National Association of Biology Teachers. She recently gave the keynote address, on "Animals and the Law" at the annual meeting of the Texas Society for Biomedical Research. drunkle@aaas.org

Acknowledgements
The project described here is funded by the Laura and John Arnold Foundation. The author thanks Michelle Barretta and Mark S. Frankel for their assistance on the paper.

References

1. The Innocence Project. www.innocenceproject.org.
2. Garret BL, quoted in http://www.businessinsider.com/its-terrifying-that-prosecutors-are-relying-on-hair-microscopy-2015-5
3. Hsu, S. FBI Admits Flaws in Hair Analysis over Decades. *Washington Post*. 2015 April 18.
4. Hsu, S. FBI Admits Flaws in Hair Analysis over Decades. *Washington Post.* 2015 April 18.
5. Editorial Board, Junk Science at the F.B.I. *The New York Times*. 2015 April 27.
6. U.S. Senate Committee on Commerce, Science, and Transportation. http://www.commerce.senate.gov/public/index.cfm?p=Hearings&ContentRecord_id=63e87410-acf3-45eb-a849-1b4edf6a8959&Statement_id=f4c8fa35-3359-4a0f-ae8c-ee2b4b0641c6&ContentType_id=14f995b9-dfa5-407a-9d35-56cc7152a7ed&Group_id=b06c39af-e033-4cba-9221-de668ca1978a&MonthDisplay=12&YearDisplay=2011.
7. Neufeld, P. The (Near) Irrelevance of *Daubert* to Criminal Justice and Some Suggestions for Reform. *Amer. J. of Public Health*. 2005; 95: S1.
8. *Daubert v. Merrell Dow Pharmaceuticals Inc.* 509 US 579 (1993).
9. Nat'l Research Council/Nat'l Academy of Science (NAS). *Strengthening Forensic Science in the United States: A Path Forward*. 2009.
10. Gianelli, P. *Daubert* and Criminal Prosecution. *Criminal Justice*. 2011, 26, 3, Fall.
11. P.L. No. 109-108, 119 Stat. 2290 (2005).
12. Charter of the Subcommittee on Forensic Science. https://www.whitehouse.gov/administration/eop/ostp/nstc.
13. White House Office of Science and Technology Policy. https://www.whitehouse.gov/sites/default/files/sofs_charter_2012_signed.pdf.
14. American Academy of Forensic Sciences. http://aafs.org/sites/default/files/2015/Non-Meeting/WHITE_HOUSE%20SUB_COMMITTEE_FS_REPORT.pdf.
15. U.S. Senate Committee on Commerce, Science, and Transportation. http://www.commerce.senate.gov/public/index.cfm?p=Hearings&ContentRecord_id=7665a46c-571e-4e73-8068-6c6402df2ae1&Statement_id=30eee171-a085-4d66-a847-

7378e94ecd7c&ContentType_id=14f995b9-dfa5-407a-9d35-56cc7152a7ed&Group_id=b06c39af-e033-4cba-9221-de668ca1978a&MonthDisplay=3&YearDisplay=2012.
16. U.S. Senate Committee on Commerce, Science, and Transportation. http://www.commerce.senate.gov/public/index.cfm?p=Hearings&ContentRecord_id=f0921154-2c02-456e-a40a-7bfa6cb417d0&Statement_id=af01eb83-8aa4-45d0-8e61-ef092eaae236&ContentType_id=14f995b9-dfa5-407a-9d35-56cc7152a7ed&Group_id=b06c39af-e033-4cba-9221-de668ca1978a&MonthDisplay=6&YearDisplay=2013.
17. Library of Congress. https://www.congress.gov/bill/112th-congress/Senate-bill/3378.
18. Charter of the U.S. Department of Justice National Commission of Forensic Science. http://www.justice.gov/sites/default/files/ncfs/legacy/2014/05/13/ncfs-charter.pdf.
19. Department of Justice-National Institute of Standards and Technology Commission on Forensic Science. http://www.justice.gov/sites/default/files/ncfs/legacy/2014/05/13/holdren-remarks.pdf.
20. American Academy of Forensic Science. www.aafs.org/sites/default/files/pdf/National%20Commission%20on%20Forensic%20Science.pdf.

A View from a member of the National Commission on Forensic Science: A Perspective on Deliberations About Forensic Science and The Path Forward

S. James Gates, Jr.

I want to begin this essay with comments on the circumstances that brought about this essay. Firstly, I owe an enormous debt to Dr. Edward Bartick, Research Professor of the Department of Forensics Sciences of George Washington University. Dr. Bartick extended the invitation for my participation in the ‘‘Forensic Science Research Evaluation Workshop’’ held at the American Association for the Advancement of Science during 26-27 May 2015. As well, I would like to acknowledge his substantial support and understanding to ensure my contribution to the proceedings of this meeting.

I also wish to pay tribute to the other participants at the meeting. Many have a deep background in the Forensic Sciences and related areas and I have greatly benefitted from this depth of knowledge as a student of the field. My own background in science is far removed from this area and accordingly what I bring to this discussion are the reflections of an outsider as the real accomplished professionals and practitioners wrestle with this exceedingly difficult challenge of finding the correct balance to ensure the products of forensic science align with the morays, standards, and traditions of fields deigned to be accurately described by the word, ‘‘science.’’

My term of service on the National Commission on Forensics Science has imbued me with a deep respect for the variegated and complicated community responsible for producing the fruits of the field to ensure *justice for all* has a firm foundation in forensic sciences and is an accurately described attribute of the U.S. system of justice. As a scientist, with over thirty years of experience in a portion of scientific disciplines extremely far removed from Forensic Science, the perspective I bring to this matter is rooted in my understanding of science across the broadest possible scales.

My charge before this gathering was to present a talk entitled, ‘‘Report on the National Commission of Sciences”[1] derived from my service on two of the Commissions subcommittees (Scientific Inquiry & Research/Training on Science & the Law). Indeed, during my talk (and below) there is a ‘‘report out’’ on the activities of the Commission. However, I feel perhaps I can add even greater value in this effort by making known the concerns I have as conceptual barriers in thinking about the work of the Commission and related activities in the U.S. Government (USG) designed to meet the challenge laid out in the landmark 2009 National Academy of Science’s report, ‘‘Strengthening Forensic Science in the United States: A Path Forward”.[2] Subsequently I will refer to this as the ‘Path Report.’

The 'Path Report' revealed a startling situation. Quoting from various parts of the report, we find

> With the exception of nuclear DNA analysis, no forensics Method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source.
>
> Some forensic science disciplines are supported by little rigorous Systematic research to validate the discipline's basic premises And techniques.
>
> Over the last two decades, advances in some forensic science disciplines, especially the use of DNA technology, have demonstrated that some areas of forensic science have great additional potential to help law enforcement identify criminals. Many crimes that may have gone unsolved are now being solved because forensic science is helping to identify the perpetrators.
>
> Those advances, however, also have revealed that, in some cases, substantive information and testimony based on faulty forensic science analyses may have contributed to wrongful convictions of innocent people.

This final sentence is especially distressing when reflecting on the final phrase of the Pledge of Allegiance of the United States. Personally, I came to understand this current status by an analogy of comparing the historical trajectory of Forensics Science to that of Medicine. Both began as noble efforts to deliver to society relief from pressing needs and arose as ''arts of practice.'' The latter, however, initiated a much earlier and more thorough uptake of the scientific method in delivering its benefits required by society. The challenge to the Forensic Science community is to now do the same.

Let's begin my comments on the National Commission on Forensic Science by briefly reviewing its structure as well as giving a report out on its work products to date.

The commission was stood up in 2013 as a joint partnership between the Department of Justice (DOJ) and the National Institute of Standards and Technology (NIST) and created to draw on the content mastery and expertise of each agency. The Commission is co-chaired by the DOJ Deputy Attorney General and the NIST Director. Currently these posts are filled by Esq. Sally Quillian Yates, and Dr. Willie May. Esq. Nelson Santos, Deputy Assistant Administrator for the Office of Forensic Sciences at the Drug Enforcement Administration, and Dr. John M. Butler, Special Assistant to the NIST Director for forensic science, serve as vice-chairs. Commission staff support is provided by Esq. Andrew J. Bruck, Senior Counsel to the Deputy Attorney General, serves as the Designated Federal Officer and Robin Jones, Consultant within the Department of Justice, serves as Program Manager, Dr. McGrath, Senior Policy Analyst with

the National Institute of Justice (NIJ), and Esq. Danielle Weiss, a Lead Associate with Booz Allen Hamilton working as a senior-level policy and strategy consultant and technical advisor to the Department of Justice's National Institute of Justice (NIJ), Office of Investigative and Forensic Sciences. There are currently twenty-four Commissioners drawn from a wide array of backgrounds including federal, state and local forensic science service providers; research scientists and academics; law enforcement officials; prosecutors, defense attorneys and judges; and other stakeholders from across the country. As well there are also eight Ex-Officio members.

The Commission holds regularly scheduled meeting open to the public and has two forms of outputs meant to influence the progress of the national forensics community to meet the challenges outlined in the 'Path Report.' One version of outputs are actually recommendations to the DOJ in the effort to put into place firm policy guidance to ensure that the Federal administration of Justice, as it relates to forensics practices and evidence, are aligned with the goals indicated as needed by the 'Path Report.' The other outputs are views documents meant to highlight for both the DOJ, as well as the larger community, sets of practices, procedures, methodologies, etc. that can serve as best practices in the difficult task of creating widespread and sustainable culture change required to meet the challenges outline in the 'Path Report.' The Commission is set up in the form of a number of subcommittees: (a.) Accreditation and Proficiency Testing, (b.) Human Factors, (c.) Interim Solutions, (d.) Medicolegal Death Investigation, (e.) Reporting and Testimony, (f.) Scientific Inquiry and Research, and (g.) Training on Science and Law. A complete listing of the output of the Commission's work can be seen on-line.

As someone from outside of the forensic science community, my reflection on the work of the Commission as a whole is it is an extreme privilege to be part of this group of accomplished dedicated and knowledgeable professionals as they wrestle with these issues. It is my hope that my perspective is useful to ensure that some out-of-box thinking is injected into the serious deliberation. It appears that the Commission is optimized in its structure to provide fruitful progress forward and has increased the depth and breadth of its work in an admirable way in service to this cause. The issue of sorting strategic versus operational level policy functions seems well underway.

Since in this workshop we have been addressing the Forensic Science research and the associated literature, I will bring to your attention the NCFS, Scientific Inquiry and Research Subcommittee's published Views document entitled "Scientific Literature in Support of Forensic Science and Practice"[1]. The following is a general statement of the NCFS, "The NCFS believes that a comprehensive evaluation of the scientific literature is critical for the advancement of forensic science policy and practice in the United States."[3]In addition, "…..it is the position of the NCFS that foundational scientific literature supportive of forensic practice should meet criteria such as the following:

- Peer-reviewed in the form of original research, substantive reviews of the original research, clinical trial reports, or reports of consensus development conferences.
- Published in a journal or book that has an International Standard Number (ISSN for journals; ISBN for books) and recognized expert(s) as authors (for books) or on its Editorial Board (for journals).
- Published in a journal that maintains a clear and publicly available statement of purpose that encourages ethical conduct such as disclosure of potential conflicts of interest integral to the peer review process
- Published in a journal that utilizes rigorous peer review with independent external reviewers to validate the accuracy in its publications and their overall consistency with scientific norms of practice.
- Published in a journal that is searchable using free, publicly available search engines (e.g. PubMed, Google Scholar, National Criminal Justice Reference Service) that search major databases of scientific literature (e.g. Medline, National Criminal Justice Reference Service Abstracts Database, and Xplore).
- Published in a journal that is indexed in databases that are available through academic libraries and other services (e.g. JSTOR, Web of Science, Academic Search Complete, and SciFinder Scholar).

While the work of the Commission is moving forward, it is also vital to keep in mind that it is only one organizational structure in an ecosystem of such organization in the USG that is tackling the challenge of the ``Path Report.'' Of very great importance is the work of the American Association for the Advancement of Science to analyze the gaps in the existing literature in assessing how large is the work necessary on a area-by-area basis that needs to be filled for there to exist an adequate scientific foundation for each one. As well within NIST, the work of the Organization of Scientific Areas Committees ''OSAC's'' is equally vital. Under the auspices of the Forensic Science Standards Board and down to the twenty-four scientific working groups (SWG's) substantial investments of energy, effort, time, and thought to solve the problems of the disciplines are occurring. It is of note that prior to the 'Path Report', the oldest SWG's date their existence back to November 1988 when the Technical Working Group for DNA Analysis Methods (TWGDAM) was formed. With its success, additional TWGs were formed for other forensic science disciplines and in 1999 the official names of the groups became Scientific Working Groups (SWGs).

I would like to close my commentary with some thoughts I regard as indicative of 'meta-challenges' to this noble effort.

As someone who has worked at the boundary of science, technology and public policy, I have come to have a keen appreciation of the role played by culture within any organized human effort. Many have observed that the culture of an organization can play an outsized role as to the effectiveness and success of an endeavor. In this regard, I have been struck by the complicated and multifaceted cultural memes that play out within the forensics science disciplines. There are at least four distinct cultural strands and models at work to my observation. These are: (a.)

engineering culture and models, (b.) law enforcement culture and models, (c.) legal culture and models, and finally (d.) scientific culture and models.

For the forensic science discipline to be successful in its delivery of products to serve the cause of justice at its highest levels, it is vital to get the mix of these four competing culture finely balanced. One of the criticisms from the 'Path Report' is that the work of forensic science can be, and in some cases was, inappropriately dominated by the influences of the law enforcement culture and hence the to call for some appropriate separation between forensics scientists and laboratories from the law enforcement agency dependent on their work.

One of the largest questions in this balance of culture issues is how closely should the standards and practices of forensic science be aligned with other areas that use the descriptor of 'science.' I have come to the opinion that the phrase 'forensic science' should not be separated in one's thinking but should be accepted as a whole in describing the activities of this community. This raises the whole question of what is the appropriate amount of science needed in the area? I do not know the answer to this question, but hope current processes have begun a care-filled and thoughtful discussion to make it so. This will be exceedingly difficult as there is a clash between four cultures surrounding this matter. The ethea of the legal, law-enforcement, forensics science and general scientific cultures must align if science is to be the ''…bulwark of criminal investigation…'' as called for by the first FBI director in 1936.[3]

There is also an amazing aspect over what I have described as ''The Dilemma Of 'Ways of Knowing.''' As I look across the boundary between science and the practice of the law, there are directly identifiable conflicts to be resolved. Some of these include the primacy of the scientific paradigm of *direct observation* versus the legal paradigm of *expert testimony*. There is the conflict between the use of the *scientific method* versus *legal precedent* to establish the 'rules of the road.' A cogent example of this can be seen in the use of the term 'scientific certainty' which most scientists would likely find a questionable belief at best and yet within the legal profession seems to possess some standing in courts of law. Finally, there is the way in which the *scientific literature* evolves versus the manner in which the body of *case law* involves. The time scales in the former are often, or even usually, far more rapid than the latter. How shall the administration of justice take this into account? How should legal precedent accommodate, or not, evolving capacities in the underlying relevant scientific areas?

There can be seen similar culture conflicts across the norms of current designations within the field of forensic science in comparison to other technical disciplines. For example, in many other technical fields, there are sharp distinctions between a scientist (usually the operative for discovery or extension of new knowledge), an engineer (the operative who creates new intellectual 'products' to translate the results of science and opens a pathway for providing practical societal benefits), and the technician (the operative who utilizes the results of engineering to actually deliver the services). From my observation, these roles are often blurred within traditional forensic science. Another point of confusion for me (perhaps fostered by my

limited exposure) is the way in which systems engineering, at its highest levels, does not seem to play a robust role in the area.

I will close this reflection by simply stating a list of questions I hope can be stimulating for those with deep mastery in the field and from which might arise new ideas in the areas.

> Since quantification of uncertainty is central to the ethos of science, how does this become resolved at the stage of judicial procedures with its varying level of proof?
>
> Should there be more elaborate structures created to coordinate the totality of responding entities, within the U. S. government, to the 2009 NAS report and other interested parties?
>
> Why are there not more opportunities for out-of-the-box ideas to come into the system (e.g. instead of chasing drugs on schedule, why not use receptor site activation binding sites patterns, etc. as a basis to determine classification of pharmaceuticals)?
>
> Is the model of 'science' even the correct conceptual framework or most urgent need to pursue? (Performance evaluation of the system as it operates currently might present an alternative goal.)

We face an enormous challenge of cultures around the foundations of science (includes its fundamental embrace of uncertainty, on-going and increasingly rapidly evolution), the role of legal precedent, good engineering practices, fair and appropriate law-enforcement engagement, and the public's and other finders of fact understanding science and standard of proof in a court of law. All of this will be needed to respond to the challenge of the 'Path Report' and create a system and society that has more fidelity to the goal of ``and justice for all.''

Sylvester James (Jim) Gates, Jr. is Distinguished University Professor, University Regents Professor and John H. Toll Professor of Physics, and the Center for Particle & String Theory Director, University of Maryland. He is a theoretical physicist. He received two B.S. degrees (mathematics & physics) in 1973 and a Ph.D. degree all from Massachusetts Institute of Technology, the latter in 1977. His doctoral thesis was the first thesis at MIT to deal with supersymmetry. He also completed postgraduate studies at both Harvard University and the California Institute of Technology (CalTech). Prof. Gates serves on the U.S. President's Council of Advisors on Science and Technology (PCAST) and the Maryland State Board of Education (MD-BoE). He serves on the National Commission of Forensic Sciences (NCFS). He is known

for his work on supersymmetry, supergravity, and superstring theory. In 1984, working with M.T. Grisaru, M. Rocek, W. Siegel, Gates co-authored *Superspace*, the first comprehensive book on the topic of supersymmetry. In 2006, he released, the book *L'arte della Fisica* (The Art of Physics), and has authored over 200 scientific publications. gatess@umd.edu

References

1. National Commission on Forensic Science. The United States Department of Justice. http://www.justice.gov/ncfs
2. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press, 2009.
3. Federal Bureau of Investigation. You Can't Get Away With It. 1936. http://www.youtube.com/watch?v=hUUgQ0ISPA8s

How Do We Trust the Scientific Literature?

Simon A. Cole

In 2009, the National Research Council report, *Strengthening Forensic Science in the United States ("Strengthening"),* stated that there was "a notable dearth of peer-reviewed, published studies establishing the scientific bases and validity of many forensic methods."[1] In response, the Research Development Technology and Evaluation Inter-agency Working Group (RDT&E IWG) of the National Science and Technology Council's Subcommittee on Forensic Science, commissioned a number of disciplinary literature compendiums, and a number of these were compiled and submitted by disciplinary groups (e.g., RDT&E IWG Hair Analysis Questions).[2,3] The National Commission on Forensic Science (NCFS), the Commission that was created as a response to *Strengthening,* then took on the issue, stating "A cursory review of the literature citations raised concerns within the NCFS that extend beyond these specific bibliographies." Under this heading, the Commission specifically identified two issues:

1. In some cases, it was unclear which literature citations are crucial to support the foundation of a particular forensic science discipline.
2. Some of the cited literature had not undergone a rigorous peer-review process.[4]

Although these statements are perhaps a bit oblique, we can probably infer what the Commission was getting at. With regard to first issue, there is a history of providing long bibliographies in support of the supposed scientific foundations of forensic disciplines. Closer inspection, in some cases, shows that the vast majority of the literature cited in these bibliographies is responsive to research questions that might be considered "not foundational." An example is latent print (fingerprint) identification, for which bibliographies primarily supporting the non-foundational question of how friction ridge skin is formed were submitted in response to legal challenges to the validity of latent print analysis.[5,6]

With regard to the second issue, it has long been alleged that much crucial information about the forensic disciplines resides in trade journals and newsletters and that these publications fall somewhat short of the Platonic ideal of a peer reviewed scientific journal, either by not engaging in peer review at all or, perhaps worse, by conducting a short of sham peer review with reviewers who either lacked scientific qualifications or who were interested in defending forensic science.[7]

There have long been claims that the forensic science journals were tainted by a guild mentality that may cause them to "avoid publishing anything that might reflect negatively on their field, thereby making forensic science journals forums for self-promotion rather than self-criticism."[8-9] For example, in the 1990s there were "allegations . . . that proponents of DNA testing have attempted to corrupt the peer-review process in order to prevent the publication of articles critical of the forensic tests."[10] Consider also the controversy over a practitioner-conducted study of confirmation bias published in *Forensic Science International (FSI),* one of

the most respected forensic journals.[11-13] On one level the controversy was about the existence of confirmation bias. But on another level, given an "astonishing" statistical error in the paper, the real question was "did anyone read this manuscript thoughtfully (before it was published)?"[14] Closer to my own specific area of interest, a recent article,[15] also in *FSI,* seems so misguided that it has caused some to wonder whether it is not merely a poor article but whether "The fact that demonstrably wrong assertions are being published in a peer-reviewed journal, such as *Forensic Science International* is a cause of concern."[15] In addition, Dror[17] reported "attempts to hide/bury/censor" the issue of cognitive and contextual bias in *The Fingerprint Sourcebook* by removing a commissioned article that had survived peer review.[18-20]

These, however, are anecdotes. Organized arguments about the deficiencies of the forensic literature are difficult to find. One effort is that Risinger and Saks.[21] Their account is unabashedly unsystematic; they characterized their examples as "canaries . . . we encountered as we labored down in our little corner of the forensic science mine, where we have for years examined reliability issues in regard to various forensic identification claims." However, they document the following issues: (1) placing some propositions beyond the reach of empirical research; (2) using research designs which cannot generate clear data on individual practitioner competence; (3) manufacturing favorable test results; (4) refusal to share data with researchers wishing to re-analyze the data; (5) encouraging overstated interpretations of data in published research reports; (6) making access to case data in FBI files contingent on accepting a member of the FBI as co-author; and (7) burying unfavorable results in reports where they are least likely to be noticed--coupled with an unexplained disclaimer that the data cannot be used to infer the false positive error rate they seem to indicate.[20]

More recently, Mnookin *et al.* noted that "Several of the most significant journals focused on publishing pattern identification research simply do not comport with broader norms of access, dissemination, or peer review typically associated with scientific publishing."[22] Especially singled out was the *AFTE Journal,* the leading disciplinary journal for firearm and toolmark analysis, which was available in very few public or university libraries and not indexed in any of the standard scientific indexes. The *Journal of Forensic Identification (JFI),* though better in these regards, also falls short as a scientific journal it in some regards; in 2008 only just over half of the Editorial Board had master's degrees or higher.[6] This situation has improved: the *JFI* now has two well credentialed scientists who serve as Technical Editors, in addition to its practitioner general editor.

Issues like these were clearly of concern to the NCFS which went on to say that "foundational scientific literature supportive of forensic practice" should meet a number of criteria: It should be subjected to independent and external peer-review; it should be published in standard academic books and journals with ISBNs or ISSNs; the editors and editorial boards should be expert in the field; and the journals should have ethics and conflict of interest codes, be free searchable, and indexed.[4] As an indicator of how things currently stand, it is worth nothing that the Scientific Inquiry and Research Subcommittee's announcement at a recent NCFS meeting that the *AFTE Journal* was now indexed in Scopus was described as "big progress."[23]

In short, the NCFS suggests that forensic science literature more faithfully emulate the familiar mainstream scientific literature. No one could dispute that these criteria are useful

minimum standards for a scientific literature. But will enforcement of these criteria allow us to trust the forensic science literature? This paper adopts a sociology of science approach to contextualize the meaning of the NCFS's critique of forensic science literature and what it means. Sociology of science treats scientific literature as a technology for conveying knowledge across time and space. As Shapin puts it, scientific literature is "a literary technology of virtual witnessing."[24] Reading scientific literature enables the reader to believe the results of a reported experiment, as though she had seen the experiment with her own eyes. This allows the research to multiply the number of people who believe in her results. "Through virtual witnessing," Shapin notes, "the multiplication of witnesses could be in principle unlimited." It is for this reason that sociologists have noted that in a certain sense the primary activity in scientific laboratories is the production scientific papers.[25] It will be noted, of course, that this is a system that depends heavily on trust (hence the title of this paper). It is only because scientists trust one another to report the results of their research truthfully that their readers are willing to read a report and behave as if they had actually seen the research performed for themselves. And, it is for this reason, of course, that breaches of that trust—scientific fraud—tend to be punished so severely in scientific communities.

Those very breaches, however, draw our attention to the fact that, even in the highest-status scientific communities, "the scientific literature" does not serve as a seamless, always-reliable conveyor of trustworthy information. Criticisms of the scientific literature and the peer review process that supposedly polices the integrity of most of it are so legion as defy recitation.[26,27] Most informed commentators today would view it as naïve to think of peer review as a powerful quality control mechanism rather than as "just as a rough-and-ready preliminary filter."[28]

While such criticisms have a long history—some say "peer review broke down as a system about forty or fifty years ago"—the situation is perhaps getting worse, rather than better.[29] High-profile frauds have continued to be exposed in natural science, medicine, and social science, of which the Michael LaCour case is only the most recent.[30] Computer scientists have written software capable of generating academic abstracts that have been accepted for presentation at major conferences.[31] The retraction of scientific papers has become so common that the phenomenon has its own dedicated web site,[32] and many disciplines and scientific institutions, including the National Science Foundation and American Association for the Advancement of Science, are concerned about crises of irreproducibility.[33] The claims described above about lack of integrity in forensic peer review don't seem out of ordinary when considering scandals in mainstream scientific peer review, including claims that peer reviews have been subpoenaed for litigation and claims of attempts by corporations to "seed" the scientific literature in anticipation of litigation.[34-36]

Moreover, recent trends in the worlds of academia and scientific research could well be exacerbating, rather than ameliorating, the problem. These include: the funding of scientific research by interested parties and conflicts of interest; accountability regimes for universities and scholars, like the British "REF" exercise; increasing budget pressures and competition for universities; the overproduction of scholars and scientists and the declining academic job market; ever-increasing standards for scholarly success (as long ago as 1992, the "most productive scientist in the world" published a paper every 3.9 days—the situation is probably even more

extreme today); the globalization of knowledge production and consumption; and library budget cuts.[37] These trends are occurring within the context of equally concerning trends in the scholarly publishing industry including: rampant multiple authoring and gift authorship; the declining size of the academic article, "salami" publishing, and "least publishable units"; the open access movement; and the erosion of the quality of peer review—possibly because the reviewers are under so much pressure to publish themselves—to the extent that many are claiming that scholarly "peer review is broken" (see Smith, this volume). A recent estimate reported that around 2 million scholarly articles are produced each year in around 25,000 scholarly journals.[38] Haack[28] observes that it was only after the Second World War that the market in scholarly publication shifted from a seller's to a buyer's market—with the explosion of post-war scientific research journals were no longer as desperate to fill their pages, and it was this that gave rise to the notion of scientific publication as a selective process. However, with the proliferation of online, open access, and even "predatory" journals,[39] the pendulum may be swinging back to a seller's market in which anyone can get practically anything published somewhere. In this context, the no-longer-recent words of Rennie[40] still resonate: "there seems to be no study too fragmented, no hypothesis too trivial, no literature citation too biased or too egotistical, no design too warped, no methodology too bungled, . . . no argument too circular, no conclusion too trifling or too unjustified, and no grammar or syntax too offensive for a paper to end up in print." In such an environment, "publication" may mean little. It is in response to such developments that critics speak of "an avalanche of low-quality research" and complain that "aspiring researchers are turned into publish-or-perish entrepreneurs."[41]

This broader context suggests that we should be cautious about assuming that the adoption of prevailing peer review and publication practices in "mainstream" science would be a panacea for whatever ills forensic science suffers. To be sure, the NCFS's suggestions enumerated above should be uncontroversial, and the forensic science enterprise would probably benefit from trying to satisfy them. In concluding, I will offer some remarks that suggest that the process of generating trust in the forensic science literature may be more complicated than simply imitating the familiar attributes of mainstream scientific literature.

First, I would note that in the discipline I know most about, latent print analysis, two recent seminal papers were not published in the forensic literature at all, but, rather, in general scientific journals. Neumann *et al.*'s paper detailing a statistical model for assessing the probative value of an association between print images was not only published in a prestigious and competitive peer-reviewed statistical journal; it was also published as a "read" paper—a sort of "super-peer review" in which 26 critical commentaries on the paper were published along with it.[42] Ulery *et al.*'s paper on a study of the accuracy of latent print reports was published in *Proceedings of the National Academy of Science,* a journal that is not only prestigious and competitive, but also viscerally appropriate, given the Academy's role in encouraging the improvement of forensic science.[43] These authors' choice of publication venue sets a exemplary model for forensic science in terms of transparency and openness to critical feedback. Perhaps these author teams merely sought the greater prestige of mainstream scientific publication. It is difficult not to suspect, however, that both teams, perhaps aware of the importance of their papers, were concerned that their results—and the peer review process—would be less well trusted if they published in a forensic specialty journal because of the suspicions about non-independence and defensiveness. And, the panel of reviewers of the Neumann *et al.* paper shows

that it was indeed reviewed by a number of professional statisticians and scientists from other fields with high statistical competence in addition to forensic practitioners. These developments prompt one to suggest that forensic science might be as well served by trying to encourage publication of important studies in mainstream scientific journals as by trying to "uplift" the forensic specialty journals.

That said, a cursory glance at a journal like *JFI* shows that, while it contains some work that might be publishable in mainstream scientific journals and some work that is mere professional news and opinion, it also contains a great deal of work, on topics like the development of latent prints on various surfaces using various visualization techniques, that is probably *not* publishable in mainstream scientific journals but nonetheless has undisputable value and utility. A venue with integrity for such useful technical work is still needed. This brings me to a second point, which is that, as I have argued elsewhere, it would be mistake to treat forensic scientists as an undifferentiated occupational group.[44] Some minimal thinking about the nature of forensic work reveals that we should realistically expect only a small portion of the people we call "forensic scientists" to be engaged in what we might call "discovery science"—the production of new knowledge about the natural world. We should expect a far greater portion to be engaged in relatively routine application of assays whose properties are relatively well understood. We should expect the needs that such groups have from "the literature" to be quite different. And, finally we might expect there to be other small groups—case managers, laboratory directors, etc.—whose needs might be different still. In sum, we need to design a forensic literature that meets the needs of "forensic scientists" as we reasonably expect them to be in order to deliver forensic services. We do not want to design a forensic literature that merely imitates the mainstream scientific literature much of which is primarily oriented toward discovery science.

Third, as Bazerman has noted, the American legal trial tends to orient around the spoken, rather than written word.[45] "The literature," therefore, tends to "enter" the trial not in its own right, but as support for the authority of some expert witness. This is problematic for a number of reasons: the court does not fully engage with the whole literature; the expert witness may have qualities, positive or negative, that distort the court's perception of the literature; writing is generally a more precise form of communication than speaking, and so on. If we are thinking about improving the forensic literature, we may want to rethink how courts (in criminal cases and otherwise) use scientific literatures. In the long run, courts may be better served by reconsidering their historic preference for the spoken word and exploring ways of engaging directly with scientific literatures themselves, rather than using expert witnesses as intermediaries (for a cautionary note, however, see Edmond).[46]

To take this point even further, though: should courts engage scientific *literature* or scientific *data*? I have already noted the irony of idealizing peer review as a solution for forensic science at a moment when peer review in mainstream science is perceived by many to be in crisis. Even more drastic, however, is the notion that the scientific article itself may be in crisis. More and more scholars and institutions are experimenting with alternative ways of communicating scientific data, realizing the obsolescence, in at least some respects, of the scientific article in the database age. Some have suggested "an end to the epiphenomenon of the text as central mediator of knowledge."[29] They are exploring, for example, web sites that display

data that can be updated in real time as further results are compiled. Especially to the extent that the mission of the forensic literature is to communicate scientific information to legal actors, like judges, jurors, and attorneys, it would be wise to think creatively about whether there are better ways than the scientific article to convey data to these actors. It is likely that few of these legal actors find the scientific article be an especially user-friendly medium in the first place. In many cases, for example, a court dealing with a live scientific controversy might be better served by a data set that is continually updated than by a static report whose obsolescence begins upon publication. (Of course, this also raises the issue of the legitimation of scientific knowledge. Some mechanism still has to serve the function traditionally served by publication—certifying to the court whether the knowledge is scientifically legitimate.) Once again, in endeavoring to improve the forensic literature, we would not want to belatedly imitate a model from which mainstream science itself has already moved on.

Simon A. Cole is Professor of Criminology, Law and Society and Director of the Newkirk Center for Science & Society at the University of California, Irvine. He received his Ph.D. in Science & Technology Studies from Cornell University. Dr. Cole is the author of *Suspect Identities: A History of Fingerprinting and Criminal Identification* (Harvard University Press, 2001), which was awarded the 2003 Rachel Carson Prize by the Society for Social Studies of Science, and more than 20 scholarly articles and book chapters about the scientific validity of fingerprint evidence and its use in the courts. His most recent book is *Truth Machine: The Contentious History of DNA Fingerprinting* (University of Chicago Press, 2008) with Michael Lynch, Ruth McNally & Kathleen Jordan. He is a recipient of a CAREER Award (2004-2009), the National Science Foundation's most prestigious award for new faculty members. He is a member of the Human Factors Subcommittee of the National Commission on Forensic Science, and he is Co-Editor of the journal *Theoretical Criminology*. scole@uci.edu

References

1. National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press, 2009.
2. Charter of the Subcommittee on Forensic Science. National Science and Technology Council. 2012. https://www.whitehouse.gov/sites/default/files/sofs_charter_2012_signed.pdf.
3. SWGFAST Response to The Research, Development, Testing & Evaluation Inter-Agency Working Group of the National Science and Technology Council, Committee on Science, Subcommittee on Forensic Science: Scientific Working Group on Friction Ridge Analysis, Study and Technology. 2011. http://www.swgfast.org/Resources/111117-ReplytoRDT&E-FINAL.pdf.
4. National Commission on Forensic Science. Scientific Literature in Support of Forensic Science and Practice. 2015. http://www.justice.gov/sites/default/files/ncfs/pages/attachments/2015/02/25/scientific_literature_views_document_as_adopted_1_30_15.pdf.
5. Cole SA. Is Fingerprint Identification Valid? Rhetorics of Reliability in Fingerprint Proponents' Discourse. Law and Policy 2006; 28(1): 109-135.
6. Cole SA. Out of the Daubert Fire and into the Fryeing Pan? The Admissibility of Latent Print Evidence in Frye Jurisdictions. Minn J Law Sci Technol 2008; 9(2):453-541

7. Benedict N. 2004. Fingerprints and the Daubert Standard for Admission of Scientific Evidence: Why Fingerprints Fail and a Proposed Remedy. Ariz Law Rev 2004; 46: 519-549.
8. Thompson WC. A Sociological Perspective on the Science of Forensic DNA Testing. U.C. Davis Law Rev 1997; 30: 1113-1136.
9. Found B. Deciphering the Human Condition: The Rise of Cognitive Forensics. Australian J Forensic Sci 2014; DOI: 10.1080/00450618.2014.965204.
10. Thompson WC. Evaluating the Admissibility of New Genetic Identification Tests: Lessons from the 'DNA War'. J Crim Law Crim 1993; 84(1): 22-104.
11. Hall LJ, Player E. Will the Instruction of an Emotional Context Affect Fingerprint Analysis and Decision Making? Forensic Sci Int 2008; 181: 36-39
12. Dror IE. On Proper Research and Understanding of the Interplay between Bias and Decision Outcomes. Forensic Sci Int 2009; 191: e17-e18.
13. Hall LJ, Player E. The Value of Practitioner Research in the Field of Fingerprint Analysis. Forensic Sci Int 2009; 191(1-3): e15-e16.
14. Saks MJ. Concerning LJ, Hall E. Player, "Will the introduction of an emotional context affect fingerprint analysis and decision-making?" Forensic Sci Int 2009; 191(1): e19.
15. Jayaprakash PT. Practical Relevance of Pattern Uniqueness in Forensic Science. Forensic Sci Int 2013; 231: 403e1-403e16.
16. Biedermann A, Curran J. Drawbacks in the Scientification of Forensic Science. Forensic Sci Int 2014; 245: e38-e40.
17. Dror IE. Cognitive Effects and Cognition of Forensic Experts. Paper read at Meeting of the National Research Council Committee on Identifying the Needs of the Forensic Sciences Community. Dec. 7, 2007, Washington, D.C.
18. Scientific Working Group on Friction Ridge Analysis Study and Technology Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST), et al., editors. The Fingerprint Sourcebook. Washington, D.C.: National Institute of Justice, 2011.
19. SWGFAST Response to The Research, Development, Testing & Evaluation Inter-Agency Working Group of the National Science and Technology Council, Committee on Science, Subcommittee on Forensic Science: Scientific Working Group on Friction Ridge Analysis, Study and Technology. 2011. http://www.swgfast.org/Resources/111117-ReplytoRDT&E-FINAL.pdf.
20. Busey TA, Dror IE. Special Abilities and Vulnerabilities in Forensic Expertise. In: The Fingerprint Sourcebook. Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) et al., editors. Washington, D.C.: National Institute of Justice, 2011.
21. Risinger D, Saks MJ. A House with No Foundation. Issues Sci Technol 2003; 20(1): 35-39
22. Mnookin JL, Cole SA, Dror IE, Fisher B, Houck MM, Inman K, Kaye DH, Koehler JJ, Langenburg G, Risinger DM, Rudin N, Siegel JA, Stoney DA. The Need for a Research Culture in the Forensic Sciences. UCLA Law Rev 2011; 58(3): 725-780.
23. Bell S. Scientific Inquiry and Research Subcommittee Report. Paper read at National Commission on Forensic Science Quarterly Meeting. May 1 2015, Washington, D.C.
24. Shapin S. Pump and Circumstance: Robert Boyle's Literary Technology. Soc Stud Sci 1984; 14: 481-520.

25. Latour B, Woolgar S. Laboratory Life: The Social Construction of Scientific Facts. Beverly Hills: Sage, 1979.
26. Chubin DE, Hackett EJ. Peerless Science: Peer Review and U.S. Science Policy. Albany: State University of New York Press, 1990.
27. Smith R. Peer Review: A Flawed Process at the Heart of Science and Medicine. J Roy Soc Med 2006; 99: 178-182.
28. Haack S. Peer Review and Publication: Lessons for Lawyers. Stetson Law Rev 2007; 36(3): 789-819.
29. Bowker GC. Emerging Configurations of Knowledge Expression. In: Media Technologies: Essays on Communication, Materiality, and Society. Gillespie T, Boczkowski PJ, Foot KA, editors. Cambridge, MA: MIT Press, 2013.
30. Bartlett T. The Unraveling of Michael LaCour. The Chronicle of Higher Education. June 2, 2015.
31. Van Noorden R. 2014. Publishers Withdraw More than 120 Gibberish Papers. Nature. Feb. 25, 2014.
32. Center for Scientific Integrity. Retraction Watch. June 9, 2015. http://retractionwatch.com/.
33. AAAS. The 'Irreproducibility' Problem. Washington, DC: American Association for the Advancement of Science, 2013.
34. Wiener J. Cancer, Chemicals and History. The Nation. Feb. 7, 2005.
35. Rosner D. Trials and Tribulations: What Happens When Historians Enter the Courtroom. Law Contemp Probl 2009; 72: 137-158
36. Freudenburg WR. Rethinking the Threats to Scientific Balance in Contexts of Litigation and Regulation. Environ Health Persp 2008; 116(1): 142-147.
37. Anderson C. Writer's Cramp. Nature 1992; 355: 101
38. Finch J. Accessibility, Sustainability, Excellence: How to Expand Access to Research Publications: U.K. Working Group on Expanding Access to Published Research Findings, 2014.
39. Bartholomew RE. Science for Sale: The Rise of Predatory Journals. J Roy Soc Med 2014; 107(10): 384-385.
40. Rennie D. Guarding the Guardians: A Conference on Editorial Peer Review. J Amer Med Assoc 1986; 256: 2391
41. Rekdal OB. 2014. Academic Urban Legends. Soc Stud Sci 2014; 44(4): 638-654.
42. Neumann C, Evett IW, Skerrett J. Quantifying the Weight of Evidence from a Forensic Fingerprint Comparison: A New Paradigm. J R Stat Soc Ser A 2012; 175(2): 371-415.
43. Ulery B, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and Reliability of Forensic Latent Fingerprint Decisions. P Natl Acad Sci USA 2011; 108(19): 7733-7738.
44. Cole SA. Acculturating Forensic Science: What Is 'Scientific Culture', and How Can Forensic Science Adopt It? Fordham Urban Law J 2010; 38(2): 435-472
45. Bazerman C. How Does Science Come to Speak in the Courts? Citations, Intertexts, Expert Witnesses, Consequential Facts, and Reasoning. Law Contemp Probl 2009; 72(1): 91-120.
46. Edmond G. Merton and the Hot Tub: Scientific Conventions and Expert Evidence in Australian Civil Procedure. Law Contemp Probl 2009; 72(1): 159-189.

Government's Role in Funding Scientific Research

Edward G. Bartick

The 2016 U.S. federal research budget was reported in *Science* to be about $84 billion.[1] The leader in research money is the National Institutes of Health (NIH) at $32 billion. Next is NASA at $19 billion, then the National Science Foundation (NSF) at $7.5 billion total budget and it goes down from there. The only federal agency whose mandated mission includes funding research in the forensic sciences is the Department of Justice (DOJ), National Institute of Justice (NIJ), Office of Investigative and Forensic Sciences (OIFS).[2] NIJ currently has a forensic science research budget of about $30 million for FY 2016, up from $22 in 2015. This budget is less than 0.1% of the NIH 2016 budget. NIJ was off the radar in the report by *Science*. The bottom line is, that with this comparatively low funding, it makes it difficult to carry out the recommendations to improve the rigor and amount of forensic science research that was recommended in the National Research Council (NRC) report of 2009[3]. My report describes NIJ's current status and discusses how other federal agencies play a role in forensic science research.

National Institute of Justice

At the NIJ the forensic and investigative sciences research priorities[4] are developed with input from the forensic science Technology Working groups[5]. The high priority needs are to improve capability to expand the information that can be extracted from traditional types of forensic evidence and to quantify the evidentiary value.

The NIJ asked the NRC to review its progress made since the 2009 report. The National Academy of Sciences released a report in December of 2015 on "Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice".[6] The report has addressed where NIJ has improved since the NRC 2009 report was released and addressed the additional improvement that is needed. The conclusion of the 2015 report's summary follows:

> The need to improve the scientific basis for some forensic disciplines is high: because of the volume of forensic transactions processed annually in the United States, even the smallest of error rates can have great consequences and erode the public's confidence in a fair and credible criminal justice system. Given NIJ's mission to serve state and local law enforcement as well as its ties to the forensic science research and practitioner communities, the agency has a unique and critical role to play in efforts to advance forensic science research.
>
> NIJ has made progress in the past 5-6 years toward improving its research operations and expanding efforts to build a research infrastructure in forensic science. Given this progress, it is now better positioned as a science agency. Although these improvements provide a solid foundation, more work is necessary to bolster NIJ's ability to advance forensic science research. This

> report offers recommendations to strengthen the role, capacity, and commitment of NIJ to support forensic science research. However, NIJ's ability to improve forensic science research in the foreseeable future will be constrained without adequate support from federal policy makers across both the executive and legislative branches. Support from DOJ leadership is particularly essential, given that NIJ's placement within the department has been perceived as a potential source of conflict of interest (NRC, 2015[6], pp. 79-80).
>
> Assuming these [the] recommendations are fully implemented and any barriers overcome, this committee believes NIJ has the potential to lead forensic science research across the federal government, a role with clear and striking consequences for the criminal justice system (NRC, 2015[6] pp.8-9).

The NRC looks forward to "a continuing enhancement of NIJ's support of forensic science research to aid in the strengthening of the analytical methods used for assisting in criminal investigation and the involvement of subjects under criminal trial".

This author's read on this NRC 2015 report is that NIJ is doing well under the circumstances. However, a real boost is needed in financial support to place the agency in a position to carry out their mission. We researchers, with an interest to work toward enhancing the rigor of forensic science, hope that this will play out. Since NIJ is not the only forensic science research supporting agency, the role other federal agencies follow.

National Science Foundation

In regard to sponsorship of extramural forensic science research, the NSF has been stimulated by the NRC 2009 report and has shown a strong interest to include forensic science research as an important funding objective. This was expressed by the initiation of its 2013 Dear Colleague letter soliciting forensic science research and education proposals.[7] Our research evaluation workshop is one outcome and NSF has embraced the partnering with NIJ on this project. A current search of NSF's awards database for active awards that mention "forensic" is about $150 million.

Additionally, the agencies have been co-sponsors to welcome proposals for the establishment of Industry/University Cooperative Research Centers (I/UCRCs) in areas relevant to the forensic sciences.[8] NSF envisions this mechanism as a means to convene experts with knowledge of the needs and challenges of forensics with vendors and developers of state-of-the-art tools, and with academic researchers at the frontiers of disciplines with the applicability to the forensic sciences. At this date, the formation of cooperatives is still in the process and eight planning grants have been awarded, totaling about $107,000, that is split 50/50 between NSF and NIJ. In the future, should a multi-site forensic science I/UCRC is launched, it could require a much larger NSF/NIJ investment over 5+ years, on the order of several million dollars, but this has not been awarded or spent yet.

The National Institute of Standards and Technology

NIST has been internally involved with forensic science for some time with a current budget of $7 million for its projects. As a result of the NRC 2009 report, its roll has been stepped up in forensic oversite. In 2013, the DOJ, partnering with NIST established the National

Commission on Forensic Science (NCFS) as an advisory means to enhance the practice and improve the reliability of forensic science.[9] Additionally, NIST has a strong oversite managing role of the Organization Scientific Area Committees (OSACs) [10,11] which have taken over the former role of Scientific Working Groups (SWGs). The OSACs have reviewed the SWG guidelines and are now setting the standard analytical methods for the forensic disciplines. Recommended research and development needs are now coming forth.[12] However, no funding commitment toward OSAC requested research has come through the federal legislature as yet. NIST is also the sponsor for the Forensic Science Center of Excellence Program, The Center for Statistics and Applications in Forensic Evidence (CSAFE)[13] has been awarded and is a consortium led by the University of Iowa and is tasked with focusing on: pattern and digital evidence with a total award of $4 million for five years to be extended depending on the success.

Federal forensic laboratories

The FBI, DEA, ATF, U.S. Post Office Laboratory, U.S. Fish and Wildlife Service Forensic Laboratory, and the Defense Forensic, Biometric Agency (DFBA), etc. conduct internal research and will sometimes provide extramural funding.[14,15] However, the extent of extramural funding provided by these agencies for forensic science research is limited and the internal research budgets are difficult to find out.

Department of Homeland Security

DHS often deals with similar technology requirements as do the federal forensic agencies and sometimes collaboratively so. The research emphasis is on explosives, chemical and biological agents and cyber technology. The investigations are more on intelligence rather than the legal aspects. The forensic agencies become involved with the justice side of the prosecution. The 2016 research budget for DHS Science & Technology is $787 million, which is less typical of law enforcement, rather more like defense intelligence. However, the Immigration Customs Enforcement (ICE) is a division of DHS and they operate the Homeland Security Investigation-Forensic Laboratory (HSI-FL) that covers explosives, drug, and other contraband. Based on the broad agency announcements, it appears that the research money is spend toward the development of border patrol methods and detection methods rather than toward the legal aspects.

National Institutes of Health and the Food and Drug Administration

The NIH and the FDA are both concerned with the toxicology of products whether they deal with medicinal, contamination of food or the improper use of drugs. They each have their own forensic chemists and do their own research, but are not likely to sponsor extramural forensic research.

Conclusion

With a very good U.S. scientific research budget in 2016, comparatively, the direct forensic science research needs are getting little attention. If the needs recommended by the NRC 2009 report are to come to fruition, a much greater commitment is necessary.

References

1. J. Mervis, Research Agencies Revel in Final 2016 Budget, *Science*, January 2016 • Vol 351 Issue 6268, pp.10-11.
2. G. LaPorte, H. Waltke and C. Heurich, National Institute of Justice Report Forensic Science, December 2015. https://www.ncjrs.gov/pdffiles1/nij/249146.pdf
3. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press, 2009.
4. NIJ Forensic Science and Investigative Sciences Research Priorities. http://www.nij.gov/topics/forensics/pages/priorities.aspx
5. NIJ Technology Working Groups. http://www.nij.gov/topics/technology/pages/working-groups.aspx
6. Committee on Strengthening Forensic Science at the National Institute of Justice, National Research Council, Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice. Washington, DC: The National Academies Press, 2015.
7. NSF 13-120, Dear Colleague Letter: Forensic Science. http://www.nsf.gov/pubs/2013/nsf13120/nsf13120.jsp
8. NSF Industry/University Cooperative Research Centers Program. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5501
9. National Commission on Forensic Science, https://www.justice.gov/ncfs
10. Organization Scientific Area Committees (OSACs). http://www.nist.gov/forensics/osac.cfm
11. M. Stolorow, Looking to the Future of Forensic Science Impacted by OSAC Standards Activities, *Forensic Science Research Evaluation*, eds., E. Bartick and M. Floyd, (2016) pp. 100-106.
12. OSAC April 2016 Newsletter, OSAC Research Need Assessment. http://www.nist.gov/forensics/osac/upload/OSAC-research-assessment-MATERIALS-Databases2016.pdf
13. 2014-NIST-FS-COE-01, Forensic Science Center of Excellence Program. http://nist.gov/forensics/center-excellence-forensic052615.cfm
14. FBI Broad Agency Announcement Counterterrorism Forensic Science R&D. www.dgmarket.com/tenders/np-notice.do?noticeId=1298447
15. Defense Forensics and Biometrics Agency: BAA. http://www.arl.army.mil/www/pages/8/DFBA%20BAA%20FY13-FY15%20approved%2016Jul2013%20updated%20links%20(2).pdf

Looking to the Future of Forensic Science Impacted by OSAC Standards Activities

Mark Stolorow

As the forensic science community is aware, national leadership in the development of a quality organizational infrastructure for the production of uniform forensic science standards and guidelines was a key element of the recommendations published in the 2009 National Academy of Sciences (NAS) report produced by the National Research Council (NRC) entitled, "Strengthening Forensic Science in the United States – A Path Forward."[1] The NRC report called for an overhaul of the forensic science community to strengthen the scientific validity of the methodologies employed by forensic science service providers, improve the quality and enforcement of standards and guidelines, develop uniformly high quality education and training curricula for both new and existing forensic science practitioners, and harmonize the practice of forensic science across more than 400 government crime laboratories in the United States. It also called for mandatory laboratory accreditation and mandatory certification of practitioners.

The NRC report specifically stated, "[t]he forensic science disciplines . . . [t]oo often have inadequate educational programs, and they typically lack mandatory and enforceable standards, founded on rigorous research and testing, certification requirements, and accreditation programs." Further, the report went on to state, "The forensic science enterprise also is hindered by its extreme disaggregation—marked by multiple types of practitioners with different levels of education and training and different professional cultures and standards for performance and a reliance on apprentice-type training and a guild-like structure of disciplines, which work against the goal of a single forensic science profession."

Specifically, the NRC Report focused on the assortment of 21 Scientific Working Groups (SWGs) that existed in 2009, variously funded by the National Institute of Justice (NIJ), the Federal Bureau of Investigation (FBI), the Drug Enforcement Agency (DEA) and other federal institutions. Although the authors of the NRC report were reasonably generous in acknowledging the laudable mission and work products of some of the various SWGs, they were highly critical of what the SWGs failed to do and the insular nature of the SWGs' structure and operations. The NRC report observed, "The efforts of these groups are laudable. However, . . . it is not clear how [they] interact or the extent to which they share requirements, standards, or policies. Thus, there is a need for more consistent and harmonized requirements."

In response to the report, the National Institute of Standards and Technology (NIST) and the U.S. Department of Justice signed a bilateral agency Memorandum of Understanding (MOU) in March 2013 which specified the establishment of the Organization of Scientific Area Committees (OSAC) model to develop standards and guidelines and promulgate implementation in all of the specific forensic science disciplines.

The consensus-based documentary standards and guidelines that will be approved for posting on the *OSAC Registry of Approved Standards* and the *OSAC Registry of Approved Guidelines* will be considered by 1) crime laboratories as standard methods for specific analyses, 2) potential discipline-specific standards for adoption as supplemental standards by accrediting bodies offering accreditation services to the forensic industry, 3) officers of the court when evaluating processes employed and testimony given by forensic science experts, and 4) by Congress and federal funding agencies in response to forensic science research priorities. The OSAC infrastructure will bring a uniform standards recognition platform to the community, enhance scientific rigor and increase communication among forensic scientists, research scientists, academicians, statisticians, attorneys, managers, quality assurance specialists and governmental entities.

How could OSAC standards affect the practice of forensic science in the laboratory and expert testimony in the courtroom in the next five to ten years? My personal aspiration is to see OSAC succeed in manifesting the following notable improvements:

1. Forensic science practitioners embrace change
2. Forensic analyses routinely employ quantification, uncertainty measurement and probabilistic modeling to the extent possible within each discipline
3. *OSAC Registry of Approved Standards* and *OSAC Registry of Approved Guidelines* become implemented in the practice of forensic science across all forensic disciplines in addition to DNA
4. Prosecutors, defense attorneys and judges begin to use the *OSAC Registry of Approved Standards* and *OSAC Registry of Approved Guidelines* in direct and cross examinations of expert witnesses
5. Judges and Juries routinely hear witnesses testify about how their analysis met current standards and scientific validity and openly describe the limitations of the tests in general and their interpretations in the specific case
6. Significant forensic science research is finally well organized and well-funded with influence from OSAC to inform research priorities
7. Standards enforcement by accrediting bodies finally becomes the rule rather than the exception.

Let's explore these seven improvements a little closer.

1. Forensic science practitioners embrace change

   Forensic science is a profession that demands of its practitioners not only excellent scientific skills but also excellent communication skills. Unlike scientists in nearly all other science professions, the forensic scientist must write laboratory reports and testify before a judge and jury in adversarial criminal court proceedings. The very nature of the adversary system encourages the lawyer who is the proponent of the introduction of forensic science expert testimony to embellish both the credentials of the forensic scientist as well as the credibility and confirmation of the proponent's version of the reconstruction of the criminal events unfavorable to the opposing side. On the other hand, it is the job of opposing counsel to impeach the forensic science expert witness, to challenge his or her credentials and the

credibility of the expert testimony. Cross examination can be and frequently is blistering and frustrating to the well-qualified forensic science expert witness. Despite the training and the knowledge he or she brings to the courtroom, remaining objective, dispassionate, balanced, cool and calm in the face of highly skilled cross examination can be very frustrating, emotionally draining and incredibly stressful.

In the face of this daunting courtroom attack, the forensic scientist must calm herself or himself and recognize that it is just another day at work. The role of the forensic science practitioner is to remain totally objective, fair, balanced and as free of bias as humanly possible under extremely stressful circumstances. The effective ones do.

However, in 2009, the forensic scientist was treated to a report by the NRC that calls into question his or her education and training, the predominance of law enforcement control and undue influence over the objective scientific decisions made by the forensic scientist in his or her analysis, conclusions, interpretation, report writing and expert testimony about the physical evidence he or she receives from the police. The report also opines that the analytical procedures, with the exception of forensic DNA testing, lacks sufficient scientific rigor underpinning the foundation of the science and lacks demonstrated validity and peer-review to pass the test of reliability. The forensic scientist is also informed by the NRC that he or she fails to control for cognitive, contextual and confirmation bias in the laboratory processes that are the hallmark of his or her profession. If the initial reaction to the criticisms in the NRC report by the forensic science community seemed a bit defensive, I hope that scholars and stakeholders can appreciate the perception by many of the forensic science practitioners that the NRC report was unduly harsh in its overall assessment in 2009, whether that assessment was accurate or not.

Well, it is my personal assessment that a lot of improvement has been made by the entire forensic science profession since 2009 and that many practitioners have accepted the reality that they must embrace change. That is what good scientists do. If forensic science practitioners are going to demonstrate to the rest of the criminal justice community that they are truly good scientists, then it is my aspiration that the current progress in embracing change will continue into the foreseeable future.

Anecdotally, one example of evidence that the forensic science community is embracing change was the successful launch of the first-ever International Symposium on Forensic Science Error Management sponsored by NIST in Washington, DC, July 21-24, 2015. Nearly 500 forensic science practitioners, managers, lawyers, judges, academic researchers, psychologists, measurement scientists and statisticians examined the detection, measurement and mitigation of errors in forensic science. Clearly, hundreds were willing to gather in recognition that forensic scientists do make errors, however infrequently, and to spend a week to examine ways to detect and manage errors in forensic science. Personally, I doubt that the concept to convene such a symposium would ever have been accepted by the forensic science community back in 2009.

2. Forensic analyses routinely employ quantification, uncertainty measurement and probabilistic modeling to the extent possible within each discipline

The NRC report was particularly critical of pattern evidence disciplines (including fingerprint, footwear and tire tracks, firearms and toolmarks, questioned documents and bite mark analysis, among others) for the subjective, qualitative nature of the examination and interpretation practices of those disciplines. It contrasted those disciplines with the objectivity and discrete features observable and quantitatively measurable in forensic DNA testing and controlled substance analysis. Consequently, it was recommended that the pattern evidence disciplines find ways to employ greater quantification, uncertainty measurement and probabilistic models.

There are currently research projects now underway in a number of forensic science disciplines to employ quantitative algorithms to objectify the data measurements of latent print minutiae and firearms and toolmark evidence striae and employ probabilistic modeling to the correlation of patterns from evidence samples and known reference samples.

It was also announced in July, 2015, by the National Institute of Standards and Technology that it would fund a new Forensic Science Center of Excellence consisting of a consortium of universities led by Iowa State University in collaboration with Carnegie Mellon University, the University of California at Irvine and the University of Virginia. The Center will be supported with $20 million in NIST funding over a five-year period and the focus of the funded research is to (1) improve the statistical foundation for pattern evidence (fingerprints, firearms, tool marks, etc.) and digital evidence (computer, video, and audio analyses, etc.) and (2) develop education and training on probabilistic methods for practitioners and other relevant stakeholders.

Clearly, there has been tangible progress made through strong national leadership in the direction of improving the employment of quantification, uncertainty measurement and probabilistic models to the extent possible within forensic science disciplines.

3. *OSAC Registry of Approved Standards* and *OSAC Registry of Approved Guidelines* become implemented in the practice of forensic science across all forensic disciplines in addition to DNA

As noted in the 2009 NRC report, the only forensic science discipline that conforms to one single set of national standard protocols is forensic DNA analysis. The reason is simple. Only forensic DNA analysis is regulated by statutory authority. The DNA Identification Act of 1994 gives the FBI Director the authority to require by law that all forensic science service providers must conform to the FBI Quality Assessment Standards (QAS) for forensic science DNA analysis in order for the DNA profiles submitted by forensic science laboratories to be accepted for uploading to the National DNA Indexing System (NDIS) via a network known as the Combined Offender DNA Indexing System (CODIS). Hundreds of crime laboratories are audited and accredited by commercial forensic science accreditation bodies in conformance with the FBI QAS for criminal offender DNA profiles and for forensic case evidence sample DNA profiles in order to be eligible to submit those DNA profiles to NDIS. Any laboratory that is not accredited to the FBI QAS standards requirements is ineligible to submit DNA profiles to NDIS. Consequently, forensic DNA analysis is the only discipline in

forensic science to have adopted one single nationally uniform set of standard protocols. In addition, all of the national forensic science accreditation bodies use the FBI QAS standards as supplemental standards to their ISO 17025 requirements.

My aspiration is to see this level of harmonization of nationally uniform sets of standard protocols become adopted and implemented by forensic science service providers and by all of the available commercial forensic science accreditation bodies, not only for forensic DNA analysis, but for all of the other forensic science disciplines as well. The pathway toward adoption and implementation will be through OSAC-facilitated development of uniform standards, posted publicly on the *OSAC Registry of Approved Standards*. Implementation into forensic science practice can then follow in two ways. First, crime laboratory management will be able to implement uniform standards by adoption from the *OSAC Registry of Approved Standards*. Second, all of the national commercial forensic science accreditation bodies have shared with OSAC their acknowledgement that they would eagerly look to the *OSAC Registry of Approved Standards* to add them to their ISO 17025 supplemental standards for the other forensic science disciplines just as they have already done for forensic DNA testing.

4. Prosecutors, defense attorneys and judges begin to use the *OSAC Registry of Approved Standards* and *OSAC Registry of Approved Guidelines* in direct and cross examinations of expert witnesses

From my personal experience in testifying as an expert in federal, state and local criminal courts, it is the rare exception rather than the rule in direct and cross examination for attorneys to question the expert about which standards he or she relied upon in forensic analysis conducted in the case at hand. It is rare for attorneys to ask the expert witness if he or she is certified as a forensic science practitioner, if his or her laboratory is accredited, or if he or she would tell the judge and jury what the currently relevant existing standards are in their forensic science discipline.

My aspiration is to see the *OSAC Registry of Approved Standards* and *Registry of Approved Guidelines* become populated with quality standards and guidelines across all of the forensic science disciplines and for those standards to become the subject of legal inquiry in examination of expert witnesses on a routine basis. Courts need to hear from the experts whether, in fact, there exist widely adopted national standards for analyzing evidence in their specific case, whether the expert applied those standards, whether there were any deviations or exceptions to those standards in his or her analysis, and what impact those deviations had on his or her conclusions. Alternatively, if national standards were not applied by the expert to the evidence in this specific case, why not?

Although, at present, one seldom hears these questions in direct and cross examination of forensic science expert witnesses, one hopes that judges and attorneys will begin to demand responses to these questions in future criminal proceedings once they become acquainted with their existence on the OSAC registries.

5. Judges and juries routinely hear witnesses testify about how their analysis met current standards and scientific validity and openly describe the limitations of the tests in general and their interpretations in the specific case

Despite the fact that a large majority of forensic science expert witnesses understand their independent role in the adversary system to deliver objective unbiased testimony, there remain conscious and unconscious stresses on all forensic science experts who anticipate the perils of cross examination. The result may often be manifested by a helpful and friendly demeanor of the expert during direct testimony but a less friendly, sometimes terse demeanor during cross examination. It requires a lot of training and dedication to duty to remain equally objective and balanced during both direct and cross examination.

Although it is understandable human nature to experience bias under stress, it is my aspiration that expert witnesses will become more helpful to judges and juries writing their reports and providing testimony in conformance to new standards that call for a description of the underlying scientific foundation supporting the validity of the testing protocols. Expert testimony might one day routinely include a description of the scientific validity of the analytical protocols used in the case as well as the limitations of the forensic conclusions for those particular analytical procedures.

6. Significant forensic science research is finally well organized and well-funded with influence from OSAC to inform research priorities

One of the criticisms of the NRC report is the shortage of research in forensic science and the reliance on non-peer reviewed publications for proof of scientific validity in some forensic disciplines. Clearly, there are relatively few sources of available funds designated specifically for forensic science research. There is also a lack of coordination across federal agencies to avoid duplication of research efforts. Perhaps the leading federal agency supporting research grants in forensic science is DOJ's National Institute of Justice (NIJ). Yet NIJ is allocated only $20-25 million annually to fund forensic science research. The Department of Defense (DOD) and the National Science Foundation (NSF) also provide funding for some forensic science related research but below the funding level of NIJ. Given that NIJ annually funds the sum of approximately $100 million for DNA rape kit backlog reduction, the amount of funding NIJ provides for forensic science research must be considered comparatively paltry.

It is my aspiration that the combined critical mass of the two new national entities, OSAC and the National Commission on Forensic Science (NCFS), will bring sufficient attention to national forensic science research needs and provide real leverage to grab the attention of lawmakers. The objectives annunciated by both organizations include defining forensic science research priorities. It is hoped that the OSAC and the NCFS articulation of these forensic science research priorities will favorably influence legislators to increase funding for research significantly. At least a tenfold increase in forensic science research funding would minimally improve the pace of strengthening the practice of forensic science but, ultimately, the effectiveness of this influence on supporting new legislation for research funding remains to be seen. One can remain hopeful.

7. Standards enforcement by accrediting bodies finally becomes the rule rather than the exception

The accrediting bodies for all of the accredited forensic science laboratories in the United States effectively use the FBI QAS standards in their audits and inspections of forensic DNA testing units. The enforcement of a single set of uniform national standards for protocols in a forensic science discipline is unique to DNA testing. The benefits are numerous. First, DNA profile comparisons are totally interoperable from all of the nearly 200 CODIS terminals currently existing in the US. Regardless of where the submitting crime laboratory is located in the U.S., a search of an unidentified questioned evidence DNA profile against more than 10 million known offender DNA profiles is achievable only because the standards required by all of the participating laboratories are uniform and interoperable.

It is my aspiration to see one day the model we observe with forensic DNA testing standards applied uniformly across all forensic science disciplines. Utilizing one set of uniform national standards for each discipline would afford improvements in the quality of testing, the confidence in test results and interoperability in comparing test results across multiple jurisdictions.

Mark D. Stolorow is the Deputy Director of the Special Programs Office (SPO) at the National Institute of Standards and Technology (NIST), U.S. Department of Commerce in Gaithersburg, Maryland where he is also the Director for OSAC Affairs (Organization of Scientific Area Committees). The mission of OSAC is to produce consensus documentary standards and guidelines to improve quality and consistency of work in the forensic science community. Mr. Stolorow has significant forensic laboratory and courtroom experience, having served as the training coordinator for the statewide forensic serology program and as the research program administrator for the Illinois State Police Bureau of Forensic Science, as well as serving as the Executive Director of Orchid Cellmark, a forensic DNA testing laboratory corporation. Mr. Stolorow led teams that performed DNA analysis for high-profile cases such as the 1995 criminal investigation of O.J. Simpson, the 1996 murder case of JonBenet Ramsey, and the 1998 case of Theodore Kaczynski. Cellmark also led the development of innovative approaches to help identify remains of 9/11 victims at the World Trade Center in New York City. Mr. Stolorow is a Fellow of the American Academy of Forensic Sciences and the recipient of the 2014 Criminalistics Section Mary E. Cowan Outstanding Service Award. He is also a member of the Midwestern Association of Forensic Scientists, ASTM and the Interagency Board of Equipment Standardization and Interoperability. Mr. Stolorow received the 2005 Midwestern Association of Forensic Scientists Distinguished Service Award. He served as co-chair of the National Science and Technology Council Subcommittee on Forensic Science from 2009 through 2012. Mr. Stolorow received a B.S. from the University of Michigan, an M.S. in Forensic Chemistry from the University of Pittsburgh, and an M.B.A. from Eastern Michigan University.
mark.stolorow@nist.gov

References

1. National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press, 2009.

Section 3: Policy Implications
Discussion Summary

The third discussion, in which policy implications were discussed, involved Ronald N. Kostoff, Deborah Runkle, Simon A. Cole, Gerald LaPorte, S. James Gates, Jr., and Mark Stolorow. This panel focused on how to move forward in the forensic sciences, and the legal repercussions of mandatory protocols and overturned decisions based on bad scientific testimony. LaPorte discussed how the National Institutes of Justice identify funding (e.g, using a Technology Working Group to identify the needs of the forensic science community), and how that funding is solicited through "Dear Colleague" letters and challenges. Cole talked about the necessity of technical studies of performance, and Kostoff stated that the forensic science community should utilize research in the basic sciences in order to avoid overlaps in particular areas of study. Stolorow discussed the enforcement of mandatory protocols by selective admittance to national databases, as well as exclusion of analysis by labs that do not follow national standards. It was agreed that the "precept of finality" of law, i.e. the permanence of rulings and decisions, has been turned on its head by science, and this may have to be considered when moving forward in the legal process.

Summary:
General Considerations for the Evaluation of Forensic Science Research Literature

The following topics and lists of their particular considerations are general considerations that should be taken into account when evaluating the Forensic Science literature. The write-ups cover the topics in more detail, but not as extensively as will be required by research on specific evidence types. For the specific evidence types, the evaluators will need in-depth knowledge of the analysis of the type being reported on. They also need an understanding of how research is conducted to produce meaningful results. The topics listed below outline the considerations required to evaluate research.

Evaluator's General Qualifications

1. Appropriate educational background
2. General understanding of the requirements of conducting research

Evaluator's Qualifications within the specific forensic discipline

1. Appropriate training
2. Appropriate experience
3. Knowledge of court requirements

Publication Category

1. Research paper
2. Technical note
3. Policy/commentary paper

Experimental Design

1. Clear hypothesis
2. Experimental design
3. Workable method in a forensic lab
4. Sufficient data
5. Use of statistics
6. Appropriate data types
7. Sufficient sample size

Statistics

1. Are statistics used?
2. Statistical ability of author(s)
3. Statistician acknowledgement
4. Use of appropriate statistics
5. Sufficient sample size

Interpretation

1. Statistically significant results

2. Reliability of the results
3. Relevance of the results
4. Court admissibility of the results

Conclusions

1. Conciseness
2. Accuracy of the interpretative points
3. Sufficient completeness of study

Bias

1. Sponsored paper - potential of conflict of interest
2. Limitations of the study?
3. Appearance of a bias?

Overall

1. Good science
2. Innovation
3. Importance
4. Usefulness
5. Confirmation of prior thought

Acknowledgements

We thank the both the National Science Foundation (NSF), Science of Science and Innovation Policy and Biological Anthropology Programs in the Directorate for Social, Behavioral and Economic Sciences (SBE), and the National Institute of Justice (NIJ), Office of Justice Programs, U.S. Department of Justice (DOJ) for sponsoring this grant. We particularly thank Maryann Feldman and Rebecca Ferrell of NSF and Gerry LaPorte of NIJ for their personal support of this project. In particular, we would like to thank Mark Weiss of the NSF (now retired) for his helpful suggestions and his inspirational encouragement for NSF's sponsorship to the make the workshop happen.

We would also like to thank the **American Association of the Advancement of Science** for providing the meeting place for this workshop. We especially thank Deborah Runkle for taking the initiative to make the meeting site available and providing many helpful suggestions for the workshop.

We thank members of the National Commission of Forensic Sciences for their review and helpful suggestions for this publication.

We are grateful for the inspiration provided by the Committee on Identifying the Needs of the Forensic Science Community, the Committee on Science, Technology, and Law Policy and Global Affairs, and the Committee on Applied and Theoretical Statistics Division on Engineering and Physical Sciences of the NATIONAL RESEARCH COUNCIL OF THE NATIONAL ACADEMIES in their two-year study and publication of *Strengthening Forensic Science in the United States, a Path Forward* (The National Academies Press, Washington, DC, (2009).

Additionally, we thank all the participants of this workshop as speakers and authors who are contributing to our effort to advance the strength of research in the forensic sciences.

We would also like to thank the planning committee for helping us to assemble a well-balanced workshop.

Workshop Planning Committee
Edward G. (Ed) Bartick, The George Washington University
Maryann Feldman, National Science Foundation
McKenzie Floyd, The George Washington University
Joseph (Jay) Kadane, Carnegie Melon University
Gerald (Gerry) Laporte, National Institute of Justice
Deborah Runkle, American Association for the Advancement of Science
Victor Weedn, The George Washington University

We would also like to especially thank Victor Weedn for his great insight and inspiration for the workshop. We spent numerous hours discussing potential speakers and topics.

Appendix

I. National Commission on Forensic Science, Scientific Inquiry and Research Subcommittee Views Document: **Scientific Literature in Support of Forensic Science and Practice**
Located on the following pages.

NATIONAL COMMISSION ON FORENSIC SCIENCE

Scientific Literature in Support of Forensic Science and Practice

Commission Action:
On January 30, 2015, the Commission voted unanimously to adopt this work product.

Type of Work Product:
Views Document issued by the Scientific Inquiry and Research Subcommittee

Statement of the Issue
It is the view of the National Commission on Forensic Science (NCFS) that each forensic discipline must have an underlying foundation that is the result of a rigorous vetting process and that is ultimately captured in the peer-reviewed scientific literature. Scientific literature comprises manuscripts that report empirical data and have been independently peer-reviewed for quality, originality, and relevance to the discipline. To strengthen confidence in results obtained in forensic examinations, each forensic discipline must identify resources that are scientifically credible, valid and with a clear scientific foundation. Such foundational literature in forensic practice should conform to norms across all scientific disciplines. Accordingly, the National Commission on Forensic Science (NCFS) proposes criteria by which scientific literature can be assessed for its consistency with principles of scientific validity.

Background
Congress called for the establishment of an independent forensic science committee at the National Academy of Sciences to assess the state of forensic sciences in the United States in 2006. The National Academy of Sciences Committee met for nearly two years (2007-2008) and summarized its findings in a report "*Strengthening Forensic Science in the United States: A Path Forward (National Research Council, 2009."* [1] During an April 23-24, 2007 committee session titled *Identifying the Needs of the Forensic Community*, the agenda included a segment on "The State of Forensic Disciplines" in which the following questions were addressed [2]: What is the state of the art? Where is research conducted? Where is it published? What is the scientific basis that informs the interpretation of the evidence? Where are advancements coming from? What are the major hurdles in the scientific foundation or methods and in the practice? What research questions would you like to have answered?" [2]

The results of this inquiry were described in the NRC report, specifically that there was "a notable dearth of peer-reviewed, published studies establishing the scientific bases and validity of many forensic methods[3]." The term "foundation" was used no less than thirty times to emphasize that each forensic discipline must have a scientifically robust and validated basis to its methods, its technologies, and its process of interpreting data.

In response to the National Research Council report regarding foundational forensic science research, an Interagency Working Group—the Research Development Technology and Evaluation (RDT&E) of the National Science and Technology Council's Subcommittee on Forensic Science was assembled and chartered with the "identification of foundational research that can be mapped to specific principles across the various disciplines of forensic science..." [4] The RDT&E committee tasked Scientific Working Groups (SWG) with addressing a series of discipline-specific questions. In response, literature compendiums were submitted to the RDT&E committee by several forensic working groups. A cursory review of the literature citations raised concerns within the NCFS that extend beyond these specific bibliographies:

1. In some cases, it was unclear which literature citations are crucial to support the foundation of a particular forensic science discipline.
2. Some of the cited literature had not undergone a rigorous peer-review process.

The goal of this Views document is to provide the framework necessary to address these and broader concerns regarding the status of the scientific foundation of forensic science across its many disciplines and practices.

View of the National Commission on Forensic Science

The NCFS believes that a comprehensive evaluation of the scientific literature is critical for the advancement of forensic science policy and practice in the United States. While other forms of dissemination of research and practice (e.g., oral and poster presentations at meetings, workshops, personal communications, editorials, dissertations, theses, and letters to editors) play an important role in science, the open, peer-reviewed literature is what endures and forms a foundation for further advancements. As stated by the National Research Council:

"Journal publication, traditionally an important means of sharing information and perspectives among scientists, is also a principal means of establishing a record of achievement in science.[5]*"*

This report further discusses the importance of issues surrounding potential conflicts of interest as it relates to peer review and publication practices:

"Disclosure, either public or institutional, is essential to controlling conflict of interest, and some universities and scientific journals prohibit certain forms of commercial contractual arrangements by their members or authors. But the responsibility for such disclosure rests with scientists themselves."[6]

Given this background and considerations, it is the position of the NCFS that foundational, scientific literature supportive of forensic practice should meet criteria such as the following:

- Peer-reviewed in the form of original research, substantive reviews of the original research, clinical trial reports, or reports of consensus development conferences.
- Published in a journal or book that has an International Standard Number (ISSN for journals; ISBN for books) and recognized expert(s) as authors (for books) or on its Editorial Board (for journals).

- Published in a journal that maintains a clear and publicly available statement of purpose that encourages ethical conduct such as disclosure of potential conflicts of interest integral to the peer review process [7]
- Published in a journal that utilizes rigorous peer review with independent external reviewers to validate the accuracy in its publications and their overall consistency with scientific norms of practice.
- Published in a journal that is searchable using free, publicly available search engines (e.g. PubMed, Google Scholar, National Criminal Justice Reference Service) that search major databases of scientific literature (e.g. Medline, National Criminal Justice Reference Service Abstracts Database, and Xplore).
- Published in a journal that is indexed in databases that are available through academic libraries and other services (e.g. JSTOR, Web of Science, Academic Search Complete, and SciFinder Scholar).

Citations

1) Strengthening Forensic Science in the United States: a Path Forward" (2009). The National Academies Press.
2) "Strengthening Forensic Science in the United States: a Path Forward" (2009). The National Academies Press. Appendix B.
3) "Strengthening Forensic Science in the United States: a Path Forward" (2009). The National Academies Press, P. 8.
4) National Science and Technology Council Committee on Science Subcommittee on Forensic Science, May 2, 2014 Office of Science & Technology Policy
5) National Research Council. "Responsibility in Science. Volume 1. Ensuring the Integrity of the Research Process" (1992) The National Academies Press, p. 52.
6) National Research Council. "Responsibility in Science. Volume 1. Ensuring the Integrity of the Research Process" (1992). The National Academies Press, p. 78-79.
7) National Research Council. "Responsibility in Science. Volume 1. Ensuring the Integrity of the Research Process" (1992). The National Academies Press, p. 141.