FORENSIC SCIENCE DISCIPLINE REVIEW OF TESTIMONY
STATISTICIAN ROUNDTABLE

Department of Justice
810 7th Street, NW
Washington, DC

July 21–22, 2016

Participants

John Butler, Ph.D., National Institute of Standards and Technology
Alicia Carriquiry, Ph.D., Iowa State University
Stephen Fienberg, Ph.D., Carnegie Mellon University
Hari Iyer, Ph.D., National Institute of Standards and Technology
Karen Kafadar, Ph.D., University of Virginia
David Kaye, M.A., J.D., Penn State Law
Steven Lund, Ph.D., National Institute of Standards and Technology
Cedric Neumann, Ph.D., South Dakota State University
Sunita Sah, Ph.D., Cornell University *[via WebEx]*
Jeff Salyards, Ph.D., Defense Forensic Science Center *[via WebEx]*
Chris Saunders, Ph.D., South Dakota State University
Hal Stern, Ph.D., University of California, Irvine

Department of Justice

Kira Antell, M.A., J.D., Office of Legal Policy
Matt Durose, M.A., Bureau of Justice Statistics
Kevin Scott, Ph.D., Office of Legal Policy
Jonathan Wroblewski, J.D., Office of Legal Policy
Victor Weedn, M.D., J.D., Office of the Deputy Attorney General

Goals and Methodology Introduction

Victor Weedn opened the meeting by welcoming the group on behalf of the Deputy Attorney General and thanking them for their participation. Dr. Weedn described forensic science as critical to the criminal justice system and to the Department as a whole.

The Office of Legal Policy (OLP) provided an introduction to the Forensic Science Discipline Review (FSDR) and the FSDR methodology development process, including a review of the time line for the FSDR. OLP explained that the objectives of the FSDR methodology development are transparency and independence. OLP then stated that the purpose of the Statistician Roundtable was to engage statisticians and researchers on assessing strengths and weaknesses of the draft methodology from statisticians' perspectives.

OLP continued by offering a brief discussion of the challenges, including the development of a standard to evaluate testimony that reconciles the concept of legal admissibility as perceived by lawyers, with statistical validity as perceived by scientists, across a five-year time period.

Selection of Cases

OLP reviewed the preliminary selection of cases that are proposed to be reviewed in the FSDR methodology – specifically, all cases from 2008 to 2012 in certain disciplines, regardless of case outcome – and opened the conversation.

Discussion:

There was significant discussion among participants about establishing what question the FSDR is attempting to answer by selecting certain cases. Participants explained that, by electing not to review cases that did not go to trial, the review will be unable to offer information on use of forensic evidence in plea discussions.

Participants generally agreed that the time frame selected was a reasonable approach to answer a narrow question about testimony in recent cases, although some expressed that a better approach would be simply to review current or ongoing cases. Two primary issues were identified with the time frame. First, participants noted that this time frame included 2009, when *Strengthening Forensic Science in the United States: A Path Forward* was published by the National Academy of Sciences.[1] Participants explained that selecting a time frame to review testimony before and after the report could reveal major differences and might answer questions as to how testimony changed as a result of the report's adopted standards. They explained, however, that one might expect non-conformities to be higher in the pre-2009 period. Second, some participants noted that the time frame was relatively arbitrary and urged the FSDR to sample cases prior to 2008. These individuals explained that sampling earlier cases would permit for greater trend analysis and further inform the results.

Participants generally felt that the disciplines proposed for review were appropriate, depending on the questions to be answered. It was suggested that the FSDR consider the most frequent

[1] The 2009 report's purpose was to highlight deficits in the practice and advancement of forensic science and to create recommendations for enforceable standards and best practices in the discipline. *See* National Academy of Science, *Strengthening Forensic Science in the United States: A Path Forward* xix (https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf). Significant changes occurred in the practice of some forensic science disciplines following the report.

activities of the lab to capture and to determine the most problematic kinds of evidence and how are they being treated.

A primary topic of discussion was the preliminary Department decision not to review FBI examiner reports in addition to testimony. Several participants strenuously advocated that the FSDR should do at least some review of reports. Participants noted that review of reports in addition to transcripts would permit an analysis of whether the reports supported the testimony – a separate and critical element. The comparison between reports and testimony could be one standard against which to evaluate conformity of testimony. Participants also noted that the reports would offer additional information about the case and variables to code, such as the types of evidence reviewed and tests performed, that ultimately may relate to testimonial outcomes.

There was also discussion about information that may not be in the reports that the FSDR should attempt to consider. In particular, there was a discussion of bias, especially as it relates to sharing of non-task relevant information with examiners. There was a suggestion that such information might not appear in a report but could influence the outcome of a trial.

An additional primary topic of discussion was the relationship between "science" and any FSDR testimonial standard. Some discussion focused on the difference between the legal standard for admissibility and "scientific validity." While OLP clarified that the FSDR is not attempting to answer the distinct question of sufficient scientific validity, participants did suggest that science played a critical role in establishing a standard. Participants urged the Department not to rely exclusively on legal admissibility as a standard. A general theme emerged as to the standard and whether the review sought to measure "correctness" or "compliance," where correctness refers to testimony based on methods that were scientifically validated, and compliance refers to testimony that was acceptable based on the prevailing view of the field as expressed in forensic science and related literature. Some participants reiterated the disconnect between legal admissibility, forensic scientist community consensus, and scientific principles, suggesting that forensic community consensus does not always rest on scientific validity.

Concern was voiced regarding the narrowness of the question to review – whether testimony given is consistent with expert consensus – and it was suggested that the FSDR needs to go beyond that to issues of scientific validity. There was a suggestion that a narrow framing (a standard tied to consensus among forensic examiners) could permit the conclusion that everything is fine and that such an assessment could be inappropriately applied more broadly. Specifically, the "narrow framing" of the question may reveal that forensic examiners complied with the consensus view in their testimonies, but that, from a scientific standpoint, such testimonies, while "compliant" and "consistent," were not scientifically justified or correct.

Participants also cautioned OLP that many people are watching this review for various reasons and that the Department must take care to define its goals carefully. Participants stressed that the Department must be careful to frame its questions, the depth of the review, and its potential outcomes precisely and narrowly.

Level of Review of Testimony

OLP discussed the proposal as to how to review the testimony and what the smallest unit of analysis could be. OLP suggested that context is relevant in such an examination, but that the goal was to preserve context and minimize judgment by the reviewer. The FSDR methodology

proposed to accomplish this objective by reviewing testimony in "threads," in which all related testimony is culled from the transcript into a thread and that thread is reviewed.

Discussion:

Participants generally felt that the threading approach was reasonable and offered advantages; however, they were unanimous that any reporting of results should use the whole testimony. Reporting at the level of the case does not preclude analysis at the level of thread or statement of relationship.

There were questions about reporting of information and whether nonconforming statements will be identified as more or less important than conforming statements, with the implication that nonconforming statements are more important because of potential justice outcomes. This led to a conversation regarding materiality, with OLP noting that a nonconforming statement – or even an identification – might not have judicial implications if the identification was not relevant to the crime at issue. OLP was cautioned not to ignore the relevance of multiple nonconforming statements because even apparently irrelevant statements could affect the outcome.

The participants also considered whether statements made during opening and closing arguments could be reviewed. While OLP shared that the goal was to review examiner statements, the participants felt that the lawyers' presentation of the forensic evidence could be an important element to review. Participants also expressed the concern that examiners' statements about their credentials and the training and proficiency processes they undergo could be important elements for review.

The group reviewed a testimony excerpt and considered and identified different statements of relationship within a worksheet model as one means to collect data. While some participants felt this was a reasonable start or approach, others felt that recording the data in rows in a worksheet could be improved by employing graphical models that allow displaying relationships between different parts of the testimony. Graphical displays might also allow for the illustration of connections between different cases (e.g., the same examiner, the same prosecutor, the same evidence type).

Standards

OLP introduced the challenge of setting the FSDR standard to be applied retrospectively – especially given the ongoing discussion about the prospective standard. OLP discussed the proposed Uniform Language for Testimony and Reports (ULTRs) project, which received over 120 comments. The ULTR comments were received from across a wide group of stakeholders and differed dramatically in opinion.

Discussion:

Initial conversation focused on the proposed ULTRs. Some participants were concerned about identifying the ULTRs as standards because standards should reflect verifiable information. Others felt the ULTRs omitted mention of error rate and other critical elements.

One option posited for a prospectively-applied standard in fields with limited foundational studies was that examiners could decline to offer opinions about the factual conclusions that the judge or jury is expected to make. Instead, the expert could simply describe the test performed and the resulting observations or data, leaving it to the trier of fact to form an opinion. It was

also noted that, with sufficient data, the expert could estimate the probability of the test results under different assumptions about the facts of the case, without giving any opinion on those facts themselves. It was echoed by some participants that the law does not require an expert to testify in the form of an opinion, and in some situations where jurors may be able to form their own conclusions and opinions without an expert opinion, this could offer a fairer way to proceed.

There was an acknowledgement that the law currently permits experts to give opinions on issues that the finder of fact must ultimately decide, and OLP indicated that, despite the wide legal latitude in allowing categorical opinions on such issues, the purpose of reviewing testimonial opinions is to assess whether the testifying expert gave an opinion with undue confidence or weight.

Significant conversation focused on the range of categorical opinions deemed acceptable for different disciplines in the ULTRs and whether these categories were helpful or implied too high a level of statistical certainty. Some suggested that no standard can have more than three categories and that they should be accompanied by statements of how often false positives and negatives occur. Among the participants who felt that opinions on the contested facts about what happened in the case should be provided, there was consensus that the number of categories should be significantly limited. As noted above, there was no unanimity as to whether experts should provide opinions, as categorization always omits information, although it does have advantages in terms of simplicity.

Significant conversation focused on the variety of outcome options for different disciplines in the ULTRs and whether these categories were helpful or implied some level of statistical certainty improperly. Some suggested that no standard could have more than three categories. While there was no unanimity as to what those three categories could be, there was a general consensus among those participants who felt that opinions were at all appropriate, that the number and scope categories should be significantly limited.

OLP listed four categories for possible standards for review for testimony in the FSDR, which could be used in some combination for the review:

1. Case reports (comparing report to testimony).
2. FSDR standard as based on the ULTRs.
3. Forensic science community consensus at the time the testimony was given.
4. Scientific literature at the time testimony was given.

Participants stressed that the correct choice of standards is driven by the research questions. More clarity is needed from OLP as to the nature of the research questions before determining the proper standard. Further, participants indicated that OLP needs to be careful about what conclusions it draws, explaining that different standards lead to different conclusions.

Scoring & Analysis

OLP introduced issues associated with evaluation of transcripts and the types of variables that would be recorded. The FSDR methodology proposes to record information on "Statements of Relationship" such as frequency, where the statement occurred, who spoke the words affirmed by the examiner, type of statement, whether the examiner improperly bolstered a statement, or whether the statement was a qualification of an earlier statement. Potential issues include varying numbers of statements in a thread, presence or absence of qualifications, quality of

qualifications, and application of differing standards. The methodology proposes to review and code each statement of relationship against the FSDR testimonial standard and, once the data have been collected, to conduct exploratory analysis.

Discussion:

The participants generally felt that exploratory review, piloting, and close analysis of transcripts prior to a full implementation would be critical. Some participants suggested beginning with a proposed list of questions and reviewing transcripts, and then repeating the process, continually refining the instrument, the questions, and the process. Participants cautioned OLP that simply beginning the process without some sort of piloting would not be successful. They explained that at the beginning, it is useful to identify specific questions or claims, or to support those claims. They cautioned that OLP should attempt to anticipate issues so there would not be a need to go back for more data.

Participants felt that in developing the protocol and the standard, it would be helpful to involve not only statisticians, but also lawyers and individuals with expertise in linguistic bias.

Participants felt that evaluation of the language used in the question in testimony was critical and questioned whether there were certain linguistic constructs (i.e. kinds of questions or phrases) from attorneys that led to more nonconformity than others. It was pointed out that without reviewing closing arguments, it may not be possible to assess whether an examiner's attempt to qualify or correct a previous statement was effective because an attorney might inaccurately reflect the findings in closing arguments. Participants were also interested in differences in testimony when examiners knew there would be an opposing expert or when there was a cross-examination. There were varying hypotheses about how this could affect testimony. If nonconformity is more or less common, it may suggest that examiners behave differently depending on circumstances.

There was an additional discussion about materiality and whether a materiality review should follow any finding of nonconformity. Participants cautioned that determining materiality is akin to determining causality and can be a very difficult task.

All participants agreed that the FSDR would be an enormous effort, and because it is always easier to collect data at one time rather than going back, they felt that the project should attempt to collect as much data and as many transcripts as possible at the outset because it would be a shame to waste, omit, or and ignore potentially revealing data due to the narrow limitations of this examination. Participants agreed that this effort could create a large database of testimony and information that could be mined by outside researchers and urged the Department to make both findings and any database created available to qualified researchers.

Reexamination of Previously Discussed Topics

OLP began by restating the FSDR purpose: to advance the use of forensic science in the courtroom by examining testimony of forensic experts in recent cases. OLP described potential outcomes to include:

- Decreasing statements in excess of a consensus standard;
- Improving examiner testimony and prosecutor questioning;
- Establishing a Department-level feedback mechanism;
- Establishing a template for reviews;

- Providing information to courtroom actors about where testimonial overstatement may occur; and
- Improving training (both internal and external).

Discussion:

There was renewed discussion of the time frame, the potential benefits of a wider sample, and the need to begin by comparing testimony to the report.

One suggestion was to make the actual statements and transcripts available – anonymized – as that would be very helpful in informing training. The idea that examiners may sometimes have been led to make nonconforming statements by attorneys seemed particularly critical for training of all parties. A standard should focus on what basis or demonstrable facts can be translated to what type of statement or testimony. Participants generally felt that a consensus standard must be involved because of the differing interpretations of what science means.

There was a discussion about the meaning of different outcomes of the FSDR. OLP was encouraged to consider what would happen if the FSDR identified no nonconformities or many nonconformities because the Department needs to consider what these outcomes would mean with respect to the stated purposes.

It was noted that one goal of the FSDR is to determine whether the Department has a pervasive problem in its recent testimony, which relates to the difference between correctness and compliance. It was posited that consensus in this study should be defined concretely; on the one hand, it might represent community-accepted practices; alternatively, it could represent what the bulk of the community reported. There was discussion of the challenges of measuring compliance when there is a desire to measure correctness (instead of compliance) by other stakeholders. There was also an acknowledgement that examiners may have complied with consensus standards in reporting a "zero error rate" or "identification to the exclusion of all others," but those statements may not be allowed today. There was a desire not to vilify examiners who were complying with policies and procedures, nor to attribute "error" to them, but participants were disinclined to wholly adopt a consensus standard – especially if that consensus standard is not statistically or scientifically validated.

Participants discussed a suggestion from a FSDR framework public commenter for a two-step review, first to compare to a current standard and then assess whether the statement could have been appropriately made using the previously-accepted standard.

There was discussion regarding what needs to occur for examiners to be able to make probabilistic statements. Participants offered that it depends on the statement the examiner wants to support. Discussion on this point focused on the kind of database that would need to be created, and what it would need to contain to make probabilistic statements about the implications of the data or simply to estimate the probabilities of the data under different hypotheses. Some suggested that, due to the difficulty of achieving a satisfactory answer on probability, presenting error rates from suitable studies along with categorical conclusions could be an alternative.

FSDR Logistics and Criteria for Expansion

OLP reviewed the proposal to: use trained administrative staff to thread testimony and remove identifiers; assign threads to trained raters for review of statements; and re-aggregate threads into testimony to permit a more complete evaluation. OLP explained that the goal was to limit human bias in all steps and permit assessment of inter-rater reliability. OLP stated that the initial implementation may be through a pilot of a single forensic science discipline that is not intended to be reviewed, or of cases outside the FSDR time period, to develop a training protocol and to ensure reliability in the process.

Discussion:

Participants generally believed this approach was reasonable but expressed concern about the structure and about the reviewers being affiliated with or employed by the Department. The proposed Departmental FSDR methodology process lacked a mechanism to step back to address these issues in a formal way. There was discussion of the kinds of models that could be used in this situation and the types of protections to be employed, including the possibility of having a data-monitoring committee. The idea of a data-monitoring committee was generally perceived as a positive interim model between full insulation within the Department and a fully independent outside review.

There was discussion about reviewing and addressing comments from stakeholder communities that may have differing opinions on how to best address issues. In particular, there was discussion about continuing to involve the forensic scientist community because the review involves them and the ultimate goal is focused on ensuring their behavior is consistent with Department efforts.

The discussion continued by reviewing the involvement of the National Commission on Forensic Science (NCFS) in the FSDR. Some indicated that the complete support of the NCFS in the FSDR was not necessary and that Commissioners were unlikely to provide the robust review the Department anticipated, due to its composition and the logistics of a 40-member committee. There was also discussion about whether some Commissioners and others in the forensic science community are too resistant to change. There was concern voiced that some Commissioners would be so resistant that they would try to block the FSDR when the participants were generally positive about the FSDR project and felt it should proceed.

Obligations and Notification

OLP introduced the potential legal or ethical need to provide notification to legal parties following identification of nonconformities.

Discussion:

There was discussion about completely insulating the FSDR from any need to report nonconformities to parties through use of a university institutional review board (IRB) model. An IRB model would set particular ground rules and identify at the outset what was shareable, what triggered a particular need to share information, and what would occur if nonconformities are discovered. This approach was contrasted to the process used in the FBI hair review, in which parties with major stakes in the outcome had voice in the process. Some participants felt

that, regardless of whether an IRB model avoiding any notification could be adopted, the Department should consider whether such a path is the correct policy.

Concluding Statements

Concluding statements generally fell into two categories: refinement of process, and methodology and standards. In addition, most statements stressed the importance of comparing testimony to examiner reports in at least some aspect of the review.

Process

- The FSDR is an enormous challenge but the approach is generally sound.
- Greater attention should be paid to the goals, with acknowledgement that the goal is not about "scientific principles."
- Independence and transparency are important, but the Department will get criticism no matter what occurs.
- Ask – what ultimate outcome is desired? Different people have different perceptions about what they want to get out of the study. Insulate it as well as possible.
- The FSDR research question should be broader.
- It is critically important to think about what the Department wants to say and to measure.
- A set of statistical principles would enhance credibility, and discussion on how to frame and insulate this process would add credibility to any forthcoming results.
- Start by thinking about what happens if one discipline or examiner is compliant in one instance, if one is non-compliant in one instance, if one is always compliant, or if one is never compliant. Then consider what you will do with the information.
- Quality assurance/quality control should be included. Flow charts and illustrations of the process come with expert human factors.
- With methodology, coding, representation, etc., there is discussion of hard science findings, but the effort to measure those elements is more an exercise in social science. Be careful not to over-quantify – this could be more of a qualitative review.
- Conduct a pilot study and after that, engage the statistical community again with preliminary findings.
- It is always challenging to review the past because science improves along the way.
- Keep lessons learned open and keep the feedback loop active.
- The unit of analysis is a holistic one related to testimony, and primary reporting has to be focused at that level.
- If there is one major change in how testimony is given, we need to be aware of that. When the study is completed, there are large-scale changes that could occur throughout the system, which should be considered in advance.
- Consider expansion. There is a cost of opportunity of not expanding the scope of work. If the scope is expanded, there is a need to rethink the period of review (2008 – 2012).
- Think about a caveat to include with the results. Various people will have various expectations and the Department needs to tell people the limitations.

Standards

- It is important to get the standards right – whether the examiners remained consistent with the range of statements within their own discipline, and consistent with the science at the time.
- The biggest concerns are authors and standards. The Department needs to limit the number of categories, at the very least. If there are categories, there must be error rates. A big chunk of forensic communication wants to improve, but in the ULTRs, we do not see that desire to improve.
- Do not conduct the FSDR until the ULTRs are agreed upon. As proposed, the ULTRs suggest a degree of precision that is not warranted given current knowledge in any forensic discipline (except single donor DNA) and should be revised.
- The perfect standard is mythical. ULTRs are critical, and feedback is useful and necessary. People will have to change, but getting them to change will be the problem. What you are trying to convey gets reported in the courtroom.
- The FSDR process is based on the existence of a standard. Aspects the standard will address include: consistent and carefully structured language; ensuring that opinions and facts are not confounded; and ensuring claims made during testimony are properly based on facts.
- The big picture is that scientific standards lead to scientific compliance (statements in the report), which leads to testimonial standards, which leads to testimonial compliance (statements in testimony).
- There are different types of compliance — the legal requirement, what the field determines, and what the scientific standard needs to address.
- The common language standard could be taken out of context. The Department needs to be very careful and make sure any measurement scales are compliant with the standards desired.

Adjournment

OLP thanked everyone for the helpful feedback provided during the two days and encouraged any participants to share any additional feedback during the formal comment period.