AN ADDENDUM TO THE PCAST REPORT ON FORENSIC SCIENCE IN CRIMINAL COURTS

On September 20, 2016, PCAST released its unanimous report to the President entitled "*Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*." This new document, approved by PCAST on January 6, 2017, is an addendum to the earlier report developed to address input received from stakeholders in the intervening period.

Background

PCAST's 2016 report addressed the question of when expert testimony based on a forensic feature-comparison method should be deemed admissible in criminal courts.[1] We briefly summarize key aspects of the previous report.

Forensic feature-comparison methods

PCAST chose to focus solely on forensic feature-comparison methods. These methods seek to determine whether a questioned sample is likely to have come from a known source based on shared features in certain types of evidence. Specific methods are defined by such elements as:

(i) the type of evidence examined (e.g., DNA, fingerprints, striations on bullets, bitemarks, footwear impressions, head-hair);
(ii) the complexity of the sample examined (e.g., a DNA sample from a single person vs. a three-person mixture in which a person of interest may have contributed only 1%); and
(iii) whether the conclusion concerns only "class characteristics" or "individual characteristics" (e.g., whether a shoeprint was made by a pair of size 12 Adidas Supernova Classic running shoes vs. whether it was made by a *specific* pair of such running shoes).

The U.S. legal system recognizes that scientific methods can assist the quest for justice, by revealing information and allowing inferences that lie beyond the experience of ordinary observers. But, precisely because the conclusions are potentially so powerful and persuasive, the law requires scientific testimony be based on methods that are scientifically valid and reliable.[2]

Requirement for empirical testing of subjective methods

In its report, PCAST noted that the *only* way to establish the scientific validity and degree of reliability of a *subjective* forensic feature-comparison method—that is, one involving significant human judgment—is to test it *empirically* by seeing how often examiners actually get the right answer. Such an empirical test of a subjective forensic feature-comparison method is referred to as a "black-box test." The point reflects a central tenet underlying all science: *an empirical claim cannot be considered scientifically valid until it has been empirically tested*.

If practitioners of a subjective forensic feature-comparison method claim that, through a procedure involving substantial human judgment, they can determine with reasonable accuracy whether a particular type of evidence came from a particular source (e.g., a specific type of pistol or a specific pistol), the claim cannot be considered scientifically valid and reliable until one has tested it by (i) providing an adequate number of examiners with an adequate number of test problems that resemble those found in forensic practice and (ii) determining whether they get the right answer with acceptable

[1] As noted in the report, PCAST did not address the use of forensic methods in criminal *investigations*, as opposed to in criminal prosecution in courts.
[2] See discussion of the Federal Rules of Evidence in Chapter 3 of PCAST's report.

frequency for the intended application.[3] While scientists may debate the precise design of a study, there is no room for debate about the absolute requirement for empirical testing.

Importantly, the test problems used in the empirical study define the specific bounds within which the validity and reliability of the method has been established (e.g., is a DNA analysis method reliable for identifying a sample that comprises only 1% of a complex mixture?).

Evaluation of empirical testing for various methods

To evaluate the empirical evidence supporting various feature-comparison methods, PCAST invited broad input from the forensic community and conducted its own extensive review. Based on this review, PCAST evaluated seven forensic feature-comparison methods to determine whether there was appropriate empirical evidence that the method met the threshold requirements of "scientific validity" and "reliability" under the Federal Rules of Evidence.

- In two cases (DNA analysis of single-source samples and simple mixtures; latent fingerprint analysis), PCAST found that there was clear empirical evidence.
- In three cases (bitemark analysis; footwear analysis; and microscopic hair comparison), PCAST found *no empirical studies whatsoever* that supported the scientific validity and reliability of the methods.
- In one case (firearms analysis), PCAST found only one empirical study that had been appropriately designed to evaluate the validity and estimate the reliability of the ability of firearms analysts to associate a piece of ammunition with a specific gun. Because scientific conclusions should be shown to be reproducible, we judged that firearms analysis currently falls short of the scientific criteria for scientific validity.
- In the remaining case (DNA analysis of complex mixtures), PCAST found that empirical studies had evaluated validity within a limited range of sample types.

Responses to the PCAST Report

Following the report's release, PCAST received input from stakeholders, expressing a wide range of opinions. Some of the commentators raised the question of whether empirical evidence is truly needed to establish the validity and degree of reliability of a forensic feature-comparison method.

The Federal Bureau of Investigation (FBI), which clearly recognizes the need for empirical evidence and has been a leader in performing empirical studies in latent-print examination, raised a different issue. Specifically, although PCAST had received detailed input on forensic methods from forensic scientists at the FBI Laboratory, the agency suggested that PCAST may have failed to take account of some relevant empirical studies. A statement issued by the Department of Justice (DOJ) on September 20, 2016 (the same day as the report's release) opined that:

> The report does not mention numerous published research studies which seem to meet PCAST's criteria for appropriately designed studies providing support for foundational validity. That omission discredits the PCAST report as a thorough evaluation of scientific validity.

Given its respect for the FBI, PCAST undertook a further review of the scientific literature and invited a variety of stakeholders—including the DOJ—to identify any "published . . . appropriately designed

[3] The size of the study (e.g., number of examiners and problems) affects the strength of conclusions that can be drawn (e.g., the upper bound on the error rate). The acceptable level of error rate depends on context.

studies" that had not been considered by PCAST and that established the validity and reliability of any of the forensic feature-comparison methods that the PCAST report found to lack such support. As noted below, DOJ ultimately concluded that it had no additional studies for PCAST to consider.

PCAST received written responses from 26 parties, including from Federal agencies, forensic-science and law-enforcement organizations, individual forensic-science practitioners, a testing service provider, and others in the US and abroad.[4] Many of the responses are extensive, detailed and thoughtful, and they cover a wide range of topics; they provide valuable contributions for advancing the field. PCAST also held several in-person and telephonic meetings with individuals involved in forensic science and law enforcement. In addition, PCAST reviewed published statements from more than a dozen forensic-science, law-enforcement and other entities.[5] PCAST is deeply grateful to all who took the time and effort to opine on this important topic.

In what follows, we focus on three key issues raised.

Issue: Are empirical studies truly necessary?

While forensic-science organizations agreed with the value of empirical tests of subjective forensic feature-comparison methods (that is, black-box tests), many suggested that the validity and reliability of such a method could be established *without* actually empirically testing the method in an appropriate setting. Notably, however, none of these respondents identified any *alternative* approach that could establish the validity and reliability of a subjective forensic feature-comparison method.

PCAST is grateful to these organizations because their thoughtful replies highlight the fundamental issue facing the forensic sciences: *the role of empirical evidence*. As noted in PCAST's report, forensic scientists rightly point to several elements that provide critical foundations for their disciplines. However, there remains confusion as to whether these elements can suffice to establish the validity and degree of reliability of particular methods.

(i) The forensic-science literature contains many papers describing variation among features. In some cases, the papers argue that patterns are "unique" (e.g., that no two fingerprints, shoes or DNA patterns are identical if one looks carefully enough). Such studies can provide a valuable *starting point* for a discipline, because they suggest that it may be worthwhile to attempt to develop reliable methods to identify the source of a sample based on feature comparison. However, such studies—no matter how extensive—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing can do so.

(ii) Forensic scientists rightly cite examiners' experience and judgment as important elements in their disciplines. PCAST has great respect for the value of examiners' experience and judgment: they are critical factors in ensuring that a scientifically valid and reliable method is practiced correctly. However, experience and judgment alone—no matter how great—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing of the method can do so.[6]

[4] www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_2016_additional_responses.pdf.
[5] www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_2016_public_comments.pdf.
[6] Some respondents, such as the Organization of Scientific Area Committees' Friction Ridge Subcommittee, suggested that forensic science should be considered as analogous to medicine, in which physicians often treat patients on the basis of experience and judgment even in the absence of established empirical evidence. However, the analogy is inapt. Physicians act with a patient's consent for the patient's benefit. There is no legal requirement, analogous to the requirement imposed upon expert testimony in court by the Federal Rules of Evidence, that physician's actions be based on "reliable principles and methods." Physicians may rely on hunches; experts testifying in court about forensic feature-comparison methods may not.

(iii) Forensic scientists cite the role of professional organizations, certification, accreditation, best-practices manuals, and training within their disciplines. PCAST recognizes that such practices play a critical role in any professional discipline. However, the existence of good professional practices alone—no matter how well crafted—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing of the method can do so.

PCAST does not diminish in any way the important roles of prior research and other types of activities within forensic science and practice. Moreover, PCAST expresses great respect for the efforts of forensic practitioners, most of whom are devoted public servants. It is important to emphasize, however, contrary to views expressed by some respondents, that there is no "hierarchy" in which empirical evidence is simply the best way to establish validity and degree of reliability of a subjective feature-comparison method. In science, empirical testing is the only way to establish the validity and degree of reliability of such an empirical method.

Fortunately, empirical testing of empirical methods is feasible. There is no justification for accepting that a method is valid and reliable in the absence of appropriate empirical evidence.

Issue: Importance of other kinds of studies

In its response to PCAST's call for further input, the Organization of Scientific Area Committees' Friction Ridge Subcommittee (OSAC FRS), whose purview includes latent-print analysis, raised a very important issue:

> While the OSAC FRS agrees with the need for black box studies to evaluate the overall validity of a particular method, the OSAC FRS is concerned this view could unintentionally stifle future research agendas aimed at dissecting the components of the black box in order to transition it from a subjective method to an objective method. If the PCAST maintains such an emphasis on black box studies as the *only* means of establishing validity, the forensic science community could be inundated with predominantly black box testing and potentially detract from progress in refining other foundational aspects of the method, such as those previously outlined by the OSAC FRS, in an effort to identify ways to emphasize objective methods over subjective methods (see www.nist.gov/topics/forensic-science/osac-research-development-needs.) Given the existing funding limitations, this will be especially problematic and the OSAC RFS is concerned other foundational research will thus be left incomplete.

PCAST applauds the work of the friction-ridge discipline, which has set an excellent example by undertaking both (i) path-breaking black-box studies to establish the validity and degree of reliability of latent-fingerprint analysis, and (ii) insightful "white-box" studies that shed light on how latent-print analysts carry out their examinations, including forthrightly identifying problems and needs for improvement. PCAST also applauds ongoing efforts to transform latent-print analysis from a subjective method to a fully objective method. In the long run, the development of objective methods is likely to increase the power, efficiency and accuracy of methods—and thus better serve the public.

In the case of subjective methods whose validity and degree of reliability have already been established by appropriate empirical studies (such as latent-print analysis), PCAST agrees that continued investment in black-box studies is likely to be less valuable than investments to develop fully objective methods. Indeed, PCAST's report calls for substantial investment in such efforts.

The situation is different, however, for subjective methods whose validity and degree of reliability has not been established by appropriate empirical studies. If a discipline wishes to offer testimony based on a subjective method, it must first establish the method's validity and degree of reliability—which can only be done through empirical studies. However, as the OSAC FRS rightly notes, a discipline could follow an alternative path by abandoning testimony based on the subjective method and instead developing an objective method. Establishing the validity and degree of reliability of an objective method is often more straightforward. PCAST agrees that, in many cases, the latter path will make more sense.

Issue: Completeness of PCAST's evaluation

Finally, we considered the important question, raised by the DOJ in September, of whether PCAST had failed to consider "numerous published research studies which seem to meet PCAST's criteria for appropriately designed studies providing support for foundational validity."

PCAST re-examined the five methods evaluated in its report for which the validity and degree of reliability had not been fully established. We considered the more than 400 papers cited by the 26 respondents; the vast majority had already been reviewed by PCAST in the course of the previous study. At the suggestion of John Butler of the National Institute of Standards and Technology (NIST), we also consulted INTERPOL's extensive summary of the forensic literature to identify additional potentially relevant papers.[7] Although our inquiry was undertaken in response to the DOJ's concern, DOJ informed PCAST in late December that it had no additional studies for PCAST to consider.

Bitemark analysis

In its report, PCAST stated that it found no empirical studies whatsoever that establish the scientific validity or degree of reliability of bitemark analysis as currently practiced. To the contrary, it found considerable literature pointing to the unreliability of the method. None of the respondents identified any empirical studies that establish the validity or reliability of bitemark analysis. (One respondent noted a paper, which had already been reviewed by PCAST, that studied whether examiners agree when measuring features in dental casts but did not study bitemarks.) One respondent shared a recent paper by a distinguished group of biomedical scientists, forensic scientists, statisticians, pathologists, medical examiners, lawyers, and others, published in November 2016, that is highly critical of bitemark analysis and is consistent with PCAST's analysis.

Footwear analysis

In its report, PCAST considered feature-comparison methods for associating a shoeprint with a specific shoe based on randomly acquired characteristics (as opposed to with a class of shoes based on class characteristics). PCAST found no empirical studies whatsoever that establish the scientific validity or reliability of the method.

The President of the International Association for Identification (IAI), Harold Ruslander, responded to PCAST's request for further input. He kindly organized a very helpful telephonic meeting with IAI member Lesley Hammer. (Hammer has conducted some of the leading research in the field—including a 2013 paper, cited by PCAST, that studied whether footwear examiners reach similar conclusions when they are presented with evidence in which the identifying features have already been identified.)

[7] The INTERPOL summaries list 4232 papers from 2010-2013 and 4891 papers from 2013-2016, sorted by discipline, see www.interpol.int/INTERPOL-expertise/Forensics/Forensic-Symposium.

Hammer confirmed that no empirical studies have been published to date that test the ability of examiners to reach correct conclusions about the source of shoeprints based on randomly acquired characteristics. Encouragingly, however, she noted that the first such empirical study is currently being undertaken at the West Virginia University. When completed and published, this study should provide the first actual empirical evidence concerning the validity of footwear examination. The types of samples and comparisons used in the study will define the bounds within which the method can be considered reliable.

Microscopic hair comparison

In its report, PCAST considered only those studies on microscopic hair comparison cited in a recent DOJ document as establishing the scientific validity and reliability of the method. PCAST found that none of these studies provided any meaningful evidence to establish the validity and degree of reliability of hair comparison as a forensic feature-comparison method. Moreover, a 2002 FBI study, by Houck and Budowle, showed that hair analysis had a stunningly high error rate in practice: Of hair samples that FBI examiners had found in the course of actual casework to be microscopically indistinguishable, 11% were found by subsequent DNA analysis to have come from different individuals.

PCAST received detailed responses from the Organization of Scientific Area Committees' Materials Subcommittee (OSAC MS) and from Sandra Koch, Fellow of the American Board of Criminalistics (Hairs and Fibers). These respondents urged PCAST not to underestimate the rich tradition of microscopic hair analysis. They emphasized that anthropologists have published many papers over the past century noting differences in average characteristics of hair among different ancestry groups, as well as variation among individuals. The studies also note intra-individual differences among hair from different sites on the head and across age.

While PCAST agrees that these empirical studies describing hair differences provide an encouraging starting point, we note that the studies do not address the validity and degree of reliability of hair comparison as a forensic feature-comparison method. What is needed are empirical studies to assess how often examiners incorrectly associate similar but distinct-source hairs (i.e., false-positive rate). Relevant to this issue, OSAC MS states: "Although we readily acknowledge that an error rate for microscopic hair comparison is not currently known, this should not be interpreted to suggest that the discipline is any less scientific." In fact, this is the central issue: the acknowledged lack of any empirical evidence about false-positive rates indeed means that, as a *forensic feature-comparison method*, hair comparison lacks a scientific foundation.

Based on these responses and its own further review of the literature beyond the studies mentioned in the DOJ document, PCAST concludes that there are no empirical studies that establish the scientific validity and estimate the reliability of hair comparison as a forensic feature-comparison method.

Firearms analysis

In its report, PCAST reviewed a substantial set of empirical studies that have been published over the past 15 years and discussed a representative subset in detail. We focused on the ability to associate ammunition not with a class of guns, but with a specific gun within the class.

The firearms discipline clearly recognizes the importance of empirical studies. However, most of these studies used flawed designs. As described in the PCAST report, "set-based" approaches can inflate examiners' performance by allowing them to take advantage of internal dependencies in the data. The

most extreme example is the "closed-set design", in which the correct source of each questioned sample is always present; studies using the closed-set design have underestimated the false-positive and inconclusive rates by more than 100-fold. This striking discrepancy seriously undermines the validity of the results and underscores the need to test methods under appropriate conditions. Other set-based designs also involve internal dependencies that provide hints to examiners, although not to the same extent as closed-set designs.

To date, there has been only one appropriately designed black-box study: a 2014 study commissioned by the Defense Forensic Science Center (DFSC) and conducted by the Ames Laboratory, which reported an upper 95% confidence bound on the false-positive rate of 2.2%.[8]

Several respondents wrote to PCAST concerning firearms analysis. None cited additional appropriately designed black-box studies similar to the recent Ames Laboratory study. Stephen Bunch, a pioneer in empirical studies of firearms analysis, provided a thoughtful and detailed response. He agreed that set-based designs are problematic due to internal dependencies, yet suggested that certain set-based studies could still shed light on the method if properly analyzed. He focused on a 2003 study that he had co-authored, which used a set-based design and tested a small number of examiners (n=8) from the FBI Laboratory's Firearms and Toolmarks Unit.[9] Although the underlying data are not readily available, Bunch offered an estimate of the number of truly independent comparisons in the study and concluded that the 95% upper confidence bound on the false-positive rate in his study was 4.3% (vs. 2.2% for the Ames Laboratory black-box study).

The Organization of Scientific Area Committee's Firearms and Toolmarks Subcommittee (OSAC FTS) took the more extreme position that all set-based designs are appropriate and that they reflect actual casework, because examiners often start their examinations by sorting sets of ammunition from a crime-scene. OSAC FTS's argument is unconvincing because (i) it fails to recognize that the results from certain set-based designs are wildly inconsistent with those from appropriately designed black-box studies, and (ii) the key conclusions presented in court do not concern the ability to sort collections of ammunition (as tested by set-based designs) but rather the ability to accurately associate ammunition with a specific gun (as tested by appropriately designed black-box studies).

Courts deciding on the admissibility of firearms analysis should consider the following scientific issues:

(i) There is only a single appropriate black-box study, employing a design that cannot provide hints to examiners. The upper confidence bound on the false-positive rate is equivalent to an error rate of 1 in 46.

(ii) A number of older studies involve the seriously flawed closed-set design, which has dramatically underestimated the error rates. These studies do not provide useful information about the actual reliability of firearms analysis.

(iii) There are several studies involving other kinds of set-based designs. These designs also involve internal dependencies that can provide hints to examiners, although not to the same extent that closed-set designs do. The large Miami-Dade study cited in the PCAST report and the small studies cited by Bunch fall into this category; these two studies have upper confidence bounds corresponding to error rates in the range of 1 in 20.

From a scientific standpoint, scientific validity should require at least two properly designed studies to ensure reproducibility. The issue for judges is whether one properly designed study, together with

[8] PCAST also noted that some studies combine tests of both class characteristics and individual characteristics, but fail to distinguish between the results for these two very different questions.

[9] PCAST did not select the paper for discussion in the report owing to its small size and set-based design, although it lists it.

ancillary evidence from imperfect studies, adequately satisfies the legal criteria for scientific validity. Whatever courts decide, it is essential that information about error rates is properly reported.

DNA analysis of complex mixtures

In its report, PCAST reviewed recent efforts to extend DNA analysis to samples containing complex mixtures. The challenge is that the DNA profiles resulting from such samples contain many alleles (depending on the number of contributors) that vary in height (depending on the ratios of the contributions), often overlap fully or partially (due to their "stutter patterns"), and may sometimes be missing (due to PCR dropout). Early efforts to interpret these profiles involved purely subjective and poorly defined methods, which were not subjected to empirical validation. Efforts then shifted to a quantitative method called combined probability of inclusion (CPI); however, this approach also proved seriously problematic.[10]

Recently, efforts have focused on an approach called probabilistic genotyping (PG), which uses mathematical models (involving a likelihood-ratio approach) and simulations to attempt to infer the likelihood that a given individual's DNA is present in the sample. PCAST found that empirical testing of PG had largely been limited to a narrow range of parameters (number and ratios of contributors). We judged that the available literature supported the validity and reliability of PG for samples with three contributors where the person of interest comprises at least 20% of the sample. Beyond this approximate range (i.e. with a larger number of contributors or where the person if interest makes a lower than 20% contribution to the sample), however, there has been little empirical validation.[11]

A recent controversy has highlighted issues with PG. In a prominent murder case in upstate New York, a judge ruled in late August (a few days before the approval of PCAST's report) that testimony based on PG was inadmissible owing to insufficient validity testing.[12] Two PG software packages (STRMix and TrueAllele), from two competing firms, reached differing[13] conclusions about whether a DNA sample in the case contained a tiny contribution (~1%) from the defendant. Disagreements between the firms have grown following the conclusion of the case.

PCAST convened a meeting with the developers of the two programs (John Buckleton and Mark Perlin), as well as John Butler from NIST, to discuss how best to establish the range in which a PG software program can be considered to be valid and reliable. Buckleton agreed that empirical testing of PG software with different kinds of mixtures was necessary and appropriate, whereas Perlin contended that empirical testing was unnecessary because it was mathematically impossible for the likelihood-ratio approach in his software to incorrectly implicate an individual. PCAST was unpersuaded by the latter argument. While likelihood ratios are a mathematically sound concept, their application requires

---

[10] Just as the PCAST report was completed, a paper was published that proposed various rules for the use of CPI. See Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M., and M.D. Coble. "Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion." *BMC Genetics*. bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7. While PCAST agreed that these rules are *necessary*, PCAST did not review whether these rules were sufficient to ensure reliability and took no position on this question.

[11] The few studies that have explored 4- or 5-person mixtures often involve mixtures that are derived from only a few sets of people (in some cases, only one). Because the nature of overlap among alleles is a key issue, it is critical to examine mixtures from various different sets of people. In addition, the studies involve few mixtures in which a sample is present at an extremely low ratio. By expanding these empirical studies, it should be possible to test validity and reliability across a broader range.

[12] See McKinley, J. "Judge Rejects DNA Test in Trial Over Garrett Phillips's Murder." New York Times, August 26, 2016, www.nytimes.com/2016/08/27/nyregion/judge-rejects-dna-test-in-trial-over-garrett-phillipss-murder.html. The defendant was subsequently acquitted.

[13] Document updated on January 17, 2017.

making a set of assumptions about DNA profiles that require empirical testing.[14] Errors in the assumptions can lead to errors in the results. To establish validity with a range of parameters, it is thus important to undertake empirical testing with a *variety* of samples in the relevant range.[15]

PCAST received thoughtful input from several respondents. Notably, one response[16] suggested that the relevant category for consideration should be expanded from "complex mixtures" (defined based on the number of contributors) to "complex samples" (defined to include also samples with low amounts of template, substantial degradation, or significant PCR inhibition, all of which will also complicate interpretation). We agree that this expansion could be useful.

The path forward is straightforward. The validity of specific PG software should be validated by testing a diverse collection of samples within well-defined ranges. The DNA analysis field contains excellent scientists who are capable of defining, executing, and analyzing such empirical studies.

When considering the admissibility of testimony about complex mixtures (or complex samples), judges should ascertain whether the published validation studies adequately address the nature of the sample being analyzed (e.g., DNA quantity and quality, number of contributors, and mixture proportion for the person of interest).

Conclusion

Forensic science is at a crossroads. There is growing recognition that the law requires that a forensic feature-comparison method be established as scientifically valid and reliable before it may be used in court and that this requirement can only be satisfied by actual empirical testing. Several forensic disciplines, such as latent-print analysis, have clearly demonstrated that actual empirical testing is feasible and can help drive improvement. A generation of forensic scientists appears ready and eager to embrace a new, empirical approach—including black-box studies, white-box studies, and technology development efforts to transform subjective methods into objective methods.

PCAST urges the forensic science community to build on its current forward momentum. PCAST is encouraged that NIST has already developed an approach, subject to availability of budget, for carrying out the functions proposed for that agency in our September report.

In addition, progress would be advanced by the creation of a cross-cutting Forensic Science Study Group—involving leading forensic and non-forensic scientists in equal measure and spanning a range of feature-comparison disciplines—to serve as a scientific forum to discuss, formulate and invite broad input on (i) empirical studies of validity and reliability and (ii) approaches for new technology development, including transforming subjective methods into objective methods. Such a forum would complement existing efforts focused on developing best practices and informing standards and might strengthen connections between forensic disciplines and other areas of science and technology. It might be organized by scientists in cooperation with one or more forensic and non-forensic science organizations—such as DFSC, NIST, IAI, and the American Association for the Advancement of Science.

[14] Butler noted that one must make assumptions, for each locus, about the precise nature of reverse and forward stutter and about the probability of allelic dropout.

[15] Butler noted that it is important to consider samples with different extents of allelic overlap among the contributors.

[16] This response was provided by Keith Inman, Norah Rudin and Kirk Lohmueller.