2023

Scientific Guidelines for Evaluating the Validity of Forensic Feature-Comparison Methods

David L. Faigman
UC Law SF, faigmand@uclawsf.edu

Nicholas Scurich

Thomas D. Albright

PNAS PERSPECTIVE OPEN ACCESS

Scientific guidelines for evaluating the validity of forensic feature-comparison methods

Nicholas Scurich[a,1], David L. Faigman[b], and Thomas D. Albright[c]

Edited by Henry Roediger III, Washington University in St. Louis, St. Louis, MO; received April 2, 2023; accepted August 17, 2023

When it comes to questions of fact in a legal context—particularly questions about measurement, association, and causality—courts should employ ordinary standards of applied science. Applied sciences generally develop along a path that proceeds from a basic scientific discovery about some natural process to the formation of a theory of how the process works and what causes it to fail, to the development of an invention intended to assess, repair, or improve the process, to the specification of predictions of the instrument's actions and, finally, empirical validation to determine that the instrument achieves the intended effect. These elements are salient and deeply embedded in the cultures of the applied sciences of medicine and engineering, both of which primarily grew from basic sciences. However, the inventions that underlie most forensic science disciplines have few roots in basic science, and they do not have sound theories to justify their predicted actions or results of empirical tests to prove that they work as advertised. Inspired by the "Bradford Hill Guidelines"—the dominant framework for causal inference in epidemiology—we set forth four guidelines that can be used to establish the validity of forensic comparison methods generally. This framework is not intended as a checklist establishing a threshold of minimum validity, as no magic formula determines when particular disciplines or hypotheses have passed a necessary threshold. We illustrate how these guidelines can be applied by considering the discipline of firearm and tool mark examination.

forensic science | Daubert | measurement | research methodology | decision-making

Forensic science has a long and storied history, dating back more than a century and is presented glowingly in classic literature and popular media alike. Statements such as the latent print found at the crime scene "matches the defendant's fingerprint" or that a bullet was fired from "the defendant's gun to the exclusion of all other guns in the world" are commonplace (1). These claims, however, have had a place not just in fiction but are the stuff of everyday courtroom testimony. One might assume that such statements are based on scientific studies that demonstrate their validity. Given the weight that fact finders might give such categorical assertions (2), this literature should be large and robust, supporting such strong claims. Unfortunately, this is not the case.

Scientists and scientific organizations are increasingly raising significant concerns about the research methods used in the limited research that has been done on forensic pattern or feature comparison methods, including fingerprints, firearms and toolmarks, bitemarks, footwear, and handwriting (e.g., refs. 3 and 4). Just on their face, forensic claims of individualization—linking a latent fingerprint to a particular person or a bullet to a specific gun—are inherently problematic (5). Outside of a vanishingly small number of areas, applied science is inherently probabilistic. Most science is directed at identifying the general phenomenon of interest, such as whether smoking can cause lung cancer, which has been referred to as an "empirical framework" (6). Whether a person's lung cancer is attributable to his smoking is a diagnostic question and is usually described in probabilistic terms based on the research exploring the empirical framework. Whether a Category 5 hurricane will hit Miami tomorrow, a person of concern will commit a violent act in the immediate future, or an eyewitness identification is accurate when it is a product of a cross-racial identification, are all hypotheses or predictions of fact that are inherently uncertain. Science gives us insights into the probability that such a hypothesis is true so that we might decide to evacuate Miami, hospitalize the person of concern, or convict the defendant.

Complicating matters, most forensic feature-comparison techniques outside of DNA are products of police laboratories rather than academic institutions of science. Nevertheless, over the decades, courts admitted these claimed areas of expertise, mainly relying on the assurances of forensic practitioners that they were valid. This practice shifted, however, with the U.S. Supreme Court's decision in Daubert v. Merrell Dow Pharmaceuticals, Inc. (7). The Daubert Court interpreted Federal Rule of Evidence 702 to require judges to examine the empirical foundation for proffered expert opinion testimony. As judges began asking about the methods, principles, and data that supported these areas of ostensible scientific evidence, they appeared to realize that little actual scientific work had been done on evidence that had long been routinely admitted (8).

Despite lacking the necessary scientific foundation, courts turned somersaults to continue admitting forensic comparison evidence in criminal trials. Courts initially ruled that Daubert's gatekeeping requirement only applied to "scientific" evidence and found most forensic areas to be "specialties," not "science."

Author affiliations: [a]Department of Psychological Science, Department of Criminology, Law and Society, University of California, Irvine, CA 92697; [b]University of California College of the Law, San Francisco, CA 94102; and [c]Salk Institute for Biological Studies, La Jolla, CA 92037

Author contributions: N.S., D.L.F., and T.D.A. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: nscurich@uci.edu.

Published October 2, 2023.

When the Supreme Court overturned this interpretation of the rule, courts still largely brought little rigor to their evaluations of non-DNA forensic evidence. Yet, over the last two decades, courts should have been alerted to the tenuous scientific foundation on which most forensic comparison fields are built. An early warning was sounded by Donald Kennedy in a 2003 Science editorial questioning the fundamental tenets of non-DNA forensic techniques (9). In addition, an extensive critique of these fields appeared in 2009 in a Report by the National Research Council (NRC), which found the following:

> With the exception of nuclear DNA analysis… no forensic 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 (3).

The President's Council of Advisors on Science and Technology (PCAST) conducted a second review of forensic comparison methods in 2016 (4) and came to similar conclusions as the 2009 NRC Report. Despite being admitted in courts in some cases for over a century, most forensic comparison methods have yet to be proven valid (10). While some courts have begun to heed these criticisms, and others have struggled to understand them, most judges continue to admit these forms of forensic evidence without serious scientific review (10–12).

This laxity appears to be a dual function of the law's inertia and ignorance of science. The inertia is largely a product of the role precedent (or "stare decisis") plays in judicial decision-making. The problem is that science operates on a fundamentally different premise. The law, by design, often perpetuates settled expectations embedded in past decisions, however science, by design, often overturns settled expectations of past research findings or beliefs. The scientific ignorance suffered by lawyers and judges is less understandable because so much of the content of modern civil and criminal cases is empirical. Daubert sought to impose on judges the responsibility for understanding the empirical grounds on which expert testimony relies and tasks lawyers appearing before them to understand the bases for the expert opinions they introduce or oppose.

The Daubert holding requiring judges to be gatekeepers against bad science is codified in Federal Rule of Evidence (FRE) 702, which most state courts largely follow. Rule 702 highlights the critical importance of empirical validation of scientific instruments, methods, and theories and requires that the expert has "reliably applied" those valid "principles and methods to the facts of the case." [FRE 702(d)] To assist trial judges in their responsibilities under Rule 702, the Daubert Court identified five nonexclusive factors that trial courts should consider when evaluating scientific evidence: Is the basis for the scientific testimony "testable," and has it been adequately tested? What is the error rate associated with the science or technique? Do adequate standards exist for the application of the method or technique? Were the findings relied on in court published in peer-reviewed journals? In addition, is the basis for the expert's testimony generally accepted in the relevant field?

While eminently reasonable—and, indeed, these factors largely reappear, explicitly or implicitly, in our discussion below—courts have had considerable difficulty employing the Daubert factors or Rule 702's standards (13). The problem is that the applied sciences appearing in court typically present these factors in highly variable ways. What adequate testing means in toxicology may not be the same as in neuroscience. The forensic pattern comparison disciplines present unique challenges regarding testing, measuring error, devising standards, providing peer review, and reaching consensus among those knowledgeable about the field. To the extent courts wish to take their gatekeeping function seriously and critically evaluate the proffered scientific evidence, they need more help than Daubert's five generic factors of sound science have so far provided.

In this Article, we set forth four guidelines that can be used to evaluate the validity of forensic feature-comparison methods designed to identify the source of a piece of forensic evidence. These guidelines are intended principally as parameters that scientists can use in designing and assessing forensic feature-comparison research. At the same time, since evidentiary rules scrutinize the science that underlies the testimony proffered in court, these guidelines should serve judicial purposes as well. Feature-comparison methods routinely involve a trained human examiner visually comparing a patterned impression left at a crime scene—fingerprints, tire tracks, firearm and toolmarks—to a known exemplar and making a subjective judgment about whether the patterns are sufficiently similar to conclude that they share a common source. In setting forth scientific guidelines for evaluating forensic comparison methods, we primarily use the domain of firearm and toolmark (FATM) identification for our discussion. This area has recently received considerable attention from researchers and courts (10, 14). It thus offers a robust example of how these guidelines might be employed to establish forensic comparison methods generally.

A Guidelines Approach for Evaluating Forensic Feature-Comparison Methods

In a highly influential article, Sir Austin Bradford Hill proposed a set of guidelines by which scientists could evaluate cause-and-effect claims in epidemiology (15). Hill's motivation was eminently practical: How could doctors identify and prevent occupational hazards from causing illness or death without first understanding which factors are causally related to sickness and injury? To that end, Hill delineated nine "aspects of association" that should be considered when evaluating causation. Although advances in technology and statistics have necessitated minor updating, the "Bradford Hill Guidelines" remain the most frequently cited and dominant framework for causal inference in epidemiology (16). Indeed, since Daubert, courts regularly rely on Hill's Guidelines in medical causation cases (17). Using Hill's approach heuristically, we propose that a guidelines approach can be of similar value for evaluating forensic pattern comparison techniques.

Inspired by Hill's framework for causal inference, we suggest the following guidelines by which to evaluate forensic pattern-comparison methods:

1. Plausibility
2. The soundness of the research design and methods: construct and external validity
3. Intersubjective testability: replication and reproducibility

4. The availability of a valid methodology to reason from group data to statements about individual cases

Close examination of these guidelines reveals that they are directed at both the conventional general or group level at which science ordinarily operates and the added question of how or whether more individualized statements about a specific source might be made. For example, in medical causation, research might well support statements such as Benzene significantly increases the population risk of leukemia; whether the research literature permits the statement that the plaintiff's leukemia was caused by Benzene exposure is a very different question (18). Forensic comparison examiners similarly claim the ability to make class-level statements—such as a bullet was shot from a Glock pistol—analogous to the group-level conclusions drawn in epidemiology. However, forensic examiners also make the much more ambitious claim that they can identify the specific source—such as a bullet was shot from the defendant's Glock. Hence, the first three guidelines are primarily directed at the empirical framework question, and the last guideline specifically considers the diagnostic question regarding making statements about individual cases.

The Guidelines. The guidelines set forth in this section are not intended as a checklist by which forensic pattern comparison fields might be evaluated. No magic formula tells researchers, much less courts, when particular disciplines or sets of hypotheses have passed a necessary threshold. Science is mainly a progressive enterprise, and our confidence, or lack of confidence, in a theory or hypothesis correlates with the amount and quality of the evidence supporting or refuting it. Therefore, none of these guidelines are, individually, necessary nor sufficient to support a conclusion regarding a forensic comparison method. Indeed, they are meant, like Hill's guidelines, to operate cumulatively, to provide more or less confidence in such opinions.

The guidelines for forensic comparisons set forth below provide four continuous measures intended to inform reasoning and decisions about the degree to which evidence proffered to a court satisfies issues regarding the plausibility of the hypotheses or theory of the field, the soundness of the research design and methods, the adequacy of the testing done, and the availability of a means to reason from group data to individual cases. Our objective is twofold. Primarily, as was the case with Hill's guidelines originally, to inform researchers seeking to validate empirical propositions that are readily amenable to test, but which are inherently probabilistic and whose "proof" is multidimensional. Secondarily, as Hill's guidelines have come to be employed, to help guide judges' decisions regarding the admissibility of forensic examiners' testimony, so that they—and the attorneys appearing before them—are provided direction about the background needed to evaluate the degree to which the evidence satisfies each guideline.

Plausibility. A fundamental starting point for all hypotheses in science concerns their basic plausibility. While it is relatively straightforward to assess the degree to which measures A and B are correlated with one another, if we are to seriously consider the hypothesis that A is a cause of B, there ordinarily exists a theory or potential mechanism to explain how that effect comes about. For instance, given what is known about physical forces, the belief that the alignment of the stars at someone's birth affects their personality is implausible. Similarly, tea leaves, crystal balls, and the lines on the palm of someone's hand are unlikely to foretell future events. In medicine, biological plausibility is inevitably considered, though sometimes plausible hypotheses turn out to be wrong—such as centuries of bleeding patients for congestive illnesses—and implausible hypotheses turn out to be correct—such as Barbara McClintock's discovery of "jumping genes" (retrotransposons), initially considered implausible but now targets of medical treatment (19, 20).

Claims in forensic pattern comparison methods that a particular fingerprint or bullet can be identified to a particular finger or gun are certainly ambitious, though not obviously implausible. Indeed, courts' faith in such claims seems to rely on little more than this plain plausibility assumption. A finger, gun, sneaker, or hammer all can leave marks that presumably contain identifying information. Matching those marks to their source seems reasonably straightforward and sensible. However, plausibility is a more complicated question than common sense alone can provide. It often involves digging deeper into the theory and methods that purportedly permit statements such as "if A then B." While fingers and guns do leave behind marks, an examiner's ability to connect marks is presumably premised on the power of the theory or method they use to accomplish this feat.

For instance, the theory of toolmark comparison developed by the Association of Firearm and Tool Mark Examiners (AFTE) allows "opinions of common origin to be made when the unique surface contours of two toolmarks are in sufficient agreement" (21). According to this theory, "sufficient agreement":

> Is related to the significant duplication of random toolmarks as evidenced by the correspondence of a pattern or combination of patterns of surface contours. Significance is determined by the comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges and furrows.... Agreement is significant when it exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool.

Finally, according to the AFTE theory, statements "that 'sufficient agreement' exists between two toolmarks means that the agreement is of a quantity and quality that the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility."

A fundamental plausibility problem arises in this theory with the standard employed to determine when marks on two items might be said to have a common source. As noted, the theory provides that "[a]greement is significant when it exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool." Hence, the AFTE theory presumes an ability among examiners to have in their heads a library of similar marks "produced by different tools" and a library of similar marks "produced by the

same tool." Judgments of identification are made by comparing the marks in question to these libraries of "similar marks" and determining that the similarities, in this case, belong in the library of same source tools rather than the library of different source tools.

The human brain, however, does not operate in this fashion. The AFTE theory contemplates that examiners behave the way a computer database might, systematically assessing the source of unknown marks to hundreds or thousands of remembered different-source marks and same-source marks and intuitively calculating the likelihood that the marks in question belong in one library or the other. This assessment leads the examiner to determine whether "agreement is of a quantity and quality that the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility." However, human brains are not supercomputers permitting this sort of systematic comparison of cases (22). While humans have modest memorization capabilities, the AFTE theory contemplates memory and analytical capacities that are implausible.

Although the AFTE theory of identification hypothesizes that the human brain functions like a supercomputer is implausible given what we know about brain function, this does not mean that pattern recognition is not viable. It may be that trained and experienced examiners develop a highly tuned intuitive ability to distinguish marks made by the same tools versus marks made by different tools. For example, experts are remarkably accurate at determining the biological sex of day-old chickens—a challenging perceptual task—even though the cognitive processes used to make these judgments are not well understood (23). Plausibility is an important consideration, but strong empirical proof could demonstrate a technique's validity even without an explanation for why or how the technique works (24).

The soundness of the research design and methods. Designing research that could—in principle—lead to strong empirical proof is a complicated matter with many different considerations and tradeoffs (25). Study design is a technical skill that benefits from various courses on methods, statistics, and measurement, and often years of advanced training and practical experience (26). Generally speaking, studies must be high in construct validity—the extent to which a test measures what it is supposed to measure—if they are to provide an empirical foundation for a technique (27). Another consideration is external validity—the extent to which the results from a study can be generalized to a population. We discuss both of these factors in relation to FATM comparison, where tests are employed to measure performance metrics such as a false positive error rate. A key lesson is that the sheer quantity of empirical studies is not a substitute for quality and that a bevy of improperly designed studies does little to provide empirical proof that the technique works.

Accurately measuring performance ("construct validity"). The standard empirical approach to establishing the validity of a method or instrument in basic and applied sciences is a "black box" study (e.g., refs. 24 and 28). Performance is evaluated as the probability that when presented with a known input, the method yields an expected output. Black box studies of forensic firearms examiners do not seek to understand how examiners reach their conclusions; they can only evaluate whether the conclusions are correct (4). For this reason, black box studies can provide only indirect corroboration of the AFTE theory; that is, such studies evaluate performance under the assumption that the comparator mechanism inside the examiner's brain is operating in accordance with the AFTE Theory (i.e., comparing the sensory difference between the present pair of known and unknown stimuli, to memorized libraries of marks produced by same tools, and memorized libraries of marks produced by different tools). Whether examiners are, in fact, following the theory as postulated by AFTE cannot be confirmed by a black box study. Nonetheless, a properly designed black box study can provide useful information about the average performance of examiners ostensibly following the AFTE theory, which may reveal examiner operating characteristics, ranging from input conditions for which examiners are proficient to those that yield performance failures.

i. Inapt design produces noisy measurements

Numerous studies of FATM identification have been conducted since the late 1990s, possibly in response to the Daubert decision requiring proof of empirical testing and error rates (29). Participants (trained forensic examiners) in these studies, ostensibly employing the FATM theory, received input in the form of multiple known and unknown bullets and were instructed to determine which unknown bullets matched the known bullets. Participants were free to make the comparisons they chose and the exact comparisons made by participants were not tracked by the researchers. These studies—which were exclusively created and conducted by FATM examiners with no specialized training in research design, statistics, and measurement—report incredible levels of human performance: Across hundreds of examiners and thousands of comparisons, zero false positive errors were made (30).

There are three basic problems with this design, commonly referred to as a "set-to-set design" (4). First, the comparisons are not independent, making statistical calculations of performance difficult, if not impossible. Second, because all of the unknowns have a corresponding known, participants could use a deductive process to reduce the number of possible matches for subsequent comparisons. Third, there were no true different-source comparisons (i.e., examiners were directly asked to evaluate two items fired by different guns), which is where a false positive error could theoretically happen. These issues make set-to-set designs inapposite for measuring examiner performance. While these studies have been presented in court by FATM examiners as precisely the empirical support that science demands (31), these fundamental design flaws are now widely recognized as precluding their ability to measure a false positive error rate. Simply put, the studies did not measure what they claimed to measure, and consequently, their results have been misrepresented in court.

Only in the last decade have FATM studies utilizing a fundamentally appropriate design been conducted. This design—known as a sample-to-sample design—gives the participant one "known" item and one "unknown" item and asks the participant to determine whether the unknown item came from the same source as the known item. The participant makes a judgment and then puts those items away. She is then presented with additional items to compare in the same

fashion. In this way, each comparison is independent, which makes calculating performance metrics relatively straightforward. To date, only five studies have utilized the sample-to-sample design, and these studies report false positive error rates of 1 to 2% (32–36). However, these studies suffer from different threats to construct validity that undermine their reported results.

ii. Existing design fails to account for a critical aspect of performance: inconclusives

Examiners following the AFTE Theory can reach three possible conclusions: identification; elimination; or inconclusive (21). Inconclusive is defined by AFTE as "an absence, insufficiency, or lack of reproducibility (of individual characteristics)" (21). Inconclusives occur in casework since evidentiary samples recovered from crime scenes may be mangled or degraded and lack markings. One survey found that FATM examiners self-report about 20% of casework being inconclusive (37). The set-to-set studies described above (e.g., ref. 30) consistently found zero or a few inconclusive responses, which one might expect given that the studies prescreened some of the test items to ensure that they had sufficient markings. However, the recent sample-to-sample studies report inconclusive responses in upward of 50 to 70% of all responses (34, 35), strongly suggesting the possibility that examiners are behaving different on a test than they do in casework. Interestingly, inconclusive responses are far more common when examiners compare bullets that were not fired by the same gun—and hence should be an elimination or exculpatory evidence—than when bullets were fired by the same gun—and hence should be an identification or inculpatory evidence (38).

How to interpret inconclusive responses on a test is a matter of significant dispute, with some arguing that they are not "errors" and others arguing that they are "potential errors" that seriously undermine the reported false positive error rates (39–44). In effect, examiners elided the difficult comparisons and aced the comparisons they selected to answer. Unfortunately, this debate about interpreting inconclusive responses cannot be resolved based on the existing studies. Therefore, even a rough estimate of the false positive error rate—a key metric for Daubert purposes—remains elusive (44). However, there are several important lessons to be learned here.

First, the recent findings on inconclusive responses demonstrate how improperly designed studies have masked a significant category of responses that could have real implications when the technique is applied in forensic contexts (45). Moreover, those studies have misrepresented performance metrics (e.g., error rates) to judges and fact-finders. Second, the existing sample-to-sample studies (e.g., refs. 32–36) are poorly designed in that roughly half of the responses are ambiguous and uninterpretable. In other scientific contexts, inconclusive can be the correct answer. For example, a medical screening test for cancer might return an ambiguous result, necessitating additional, possibly more expensive and intrusive tests. FATM studies could follow suit and include by design ambiguous comparisons for which inconclusive is the correct response (39, 46). Statistical techniques have been developed to analyze expert performance in which inconclusives or multiple other categories exist (46). Third, the existing sample-to-sample studies reveal an important characteristic of how FATM examiners operate: They are very unlikely to call inconclusive for inculpatory evidence and very likely to call inconclusive for exculpatory evidence (38). This asymmetry is not contemplated in the AFTE protocol and therefore reveals a significant discrepancy between the AFTE theory and how it is effectuated by practicing FATM examiners.

iii. Lack of control group

The usual practice in research is to create comparison groups, often referred to as experimental and control groups, in order to measure the performance or effects of some variable of interest. Indeed, the Food and Drug Administration typically will not approve drugs for commercial use unless the validation study contains a (placebo) control group to which the treatment group is compared (47). The threshold question is whether the drug produces its intended effect to a greater extent in the experimental group, rather than the control group.

The analog to treatment and control groups in the forensic pattern comparison domains is comparing the performance of experts to novices. In science, expertise is defined as superior performance within the particular domain of claimed expertise (48). Demonstration of superior performance requires direct comparison of experts to novices. This feature is glaringly absent in FATM research. Without suitable controls, these studies provide no information regarding the difficulty of the tasks presented, differences between groups of examiners based on level of experience and training, differences in performance rates between alternative methods or protocols of identification, and so forth (49). Indeed, a group of untrained defense attorneys passed one study "with flying colors"—despite the fact that the study had been proffered in multiple court proceedings as evidence validating the FATM field (50).

Some pattern comparison methods have included control groups in their studies, though these studies have tended to be small in scale and rare (51). Beyond simply comparing performance, these studies provide important insights into the decision-making processes of experts, such as how those processes might differ from novices (e.g., more conservative in calling identifications), and what types of comparisons might require more training or represent a hard limit on the capacities of human decision makers (52).

Generalizing research findings to fieldwork ("external validity"). A fundamental goal in scientific research is to elucidate general principles that have the power to make accurate predictions across a broad range of conditions in the real world. Experimentation conducted in a laboratory has multiple benefits, including tight control over extraneous variables and knowledge of ground truth of the comparisons presented to participants in black box studies. However, these benefits come at a cost: Concerns are frequently expressed about the artificiality of the laboratory simulation or the extent to which results from a sample of participants apply to a broader population.

The ability to generalize laboratory findings is known as "external validity" (53). External validity is an empirical question that must be established through testing. Basic scientific

studies reveal the operating characteristics of vision—sensitivity, discriminability, and recognition ability—have demonstrated external validity well beyond the conditions of the experiments in which they were revealed (54–56). However, the external validity of FATM black box studies has not been tested and has been challenged on two grounds: the lack of correspondence between the research studies and fieldwork and how the participants were sampled. Both factors have led commentators to argue against presuming the results of existing black box studies apply to the field writ large.

i. Important differences between research and fieldwork

In research studies, examiners know they are participating in a study designed to measure their performance, and they are instructed to work alone and not to consult peers or have their work reviewed by peers. This differs from casework in that accredited laboratories require a "peer review" or verification by a second examiner. How this second review might impact the results of a FATM examination is unknown. For example, some have argued that errors committed in a research study would have been detected in casework by the second examiner reviewing the first examiner's work (32, 57). Therefore, the error rates in casework would be lower than the error rates reported in studies. This conjecture is ultimately an empirical question, and it critically depends on the second examiner being blinded to the conclusion reached by the first examiner—which is quite uncommon in laboratories and not required by accrediting bodies (58). Not surprisingly, then, evidence indicates that it is very rare for examiners to disagree with one another in casework (37).

Another concern is that examiners may behave differently under testing conditions than they would in casework. The fact that 50 to 70% of study responses are inconclusive while the rate of inconclusives in casework is considerably less suggests that something is afoot. Specifically, examiners partaking in a study may alter their thresholds for identification decisions relative to the thresholds employed in laboratory studies (59, 60). As a result, examiners may be more likely to call inconclusive on a test, than in casework, since inconclusive responses occasion no negative consequences.

Recognition of this issue has motivated the use of "blind proficiency testing," where sample items are filtered into casework unbeknownst to examiners (61). Koehler has appropriately been advocating for blind proficiency testing for over a decade, unfortunately to little effect (62). A notable exception has been a testing regime conducted at the Houston Forensic Science Center, which regularly includes test items in casework to ensure laboratory procedures are followed (63). This testing regime is a strong counter to the oft-made claims that blind testing is too costly, impractical, or even impossible. While costly in terms of time and resources needed to defeat examiner awareness of being tested (64), this testing method has far greater external validity for estimates of examiner performance on actual casework (42).

ii. Sample selection effects

Laboratory studies rely on a sample of examiners in the hope that the findings from the sample can be extrapolated to the broader population of examiners. Unfortunately, studying the entire population of FATM examiners is infeasible, so studying a sample of examiners is necessary. The ability to extrapolate the findings of a sample to the broader population critically depends on several factors related to how the participants are selected and whether they drop out of the study or fail to complete items randomly (44).

Before one even considers drawing a sample from a population of examiners, the population needs to be clearly defined. It might, for example, consist of all registered voters in the United States in 2016 if one were interested in predicting who might be elected President of the United States. However, in the FATM domain, would this be all FATM examiners? Only those accredited by AFTE? Only those who regularly testify in criminal trials? etc. What's more, the size of these populations is unknown. For example, it has been estimated that AFTE has approximately 1,200 members, though a sizeable portion is thought to be inactive and lacks a valid email address. This lack of information makes it impossible to assess whether the demographics of a given sample mirror the population. It also precludes the use of probability sampling methods—the gold standard for collecting representative samples in science (65).

All of the FATM studies rely on volunteers. Research across various domains finds that volunteers willing to participate in research perform differently than nonvolunteers (66). Using volunteers creates the danger that performance among research subjects is better (or different) than what could be expected from typical examiners. If those with high confidence, and potentially better skills, are subjects in the research, the results provide no insights regarding the average performance of examiners in the field. Indeed, Koehler noted, "A testing programme that relies on voluntary participation will not produce trustworthy data because the sampled population may no longer be representative of testifying examiners (62 at p. 94)."

An interesting aspect of the concern over representativeness is the question of whether the experience and training of an examiner affect performance. A cornerstone principle of FATM practice—and most all forensic pattern comparison methods—is that experience and training give examiners the expertise to distinguish same-source from different-source markings. However, in one study reporting on the relationship between experience and performance, none was found (67). This should have surprised the researchers, given findings on perceptual learning in general (68), but was largely left unexplored. This issue suggests a critical line of inquiry for future research and a possible variable on which decisions about the admissibility of FATM experts might be based (48).

Another significant concern in FATM research is attrition bias, which occurs when participants drop out of the study after initiating it. Analyzing only the results of the participants who chose to remain in the study can lead to biased statistical estimates and faulty conclusions (69). As a rule of thumb, one group of medical researchers posited that concern about bias is warranted when 20% of participants drop out (70); one study of FATM examiners reported that 69% of the participants who started the study failed to complete the entire study (67). Many FATM studies do not bother to report dropout rates (36).

An interesting remedy to the problems of relying on volunteers and attrition bias—other than surreptitiously including test items in casework—is to require all examiners to complete

the study. Koehler and Liu used this approach in the domain of latent fingerprint examination and found error rates significantly higher than those in studies that rely on volunteers (71), which corroborates concerns about crediting the results of studies that rely on convenience samples of volunteers and that have an enormous amount of attrition.

Intersubjective testability: Replication and reproducibility. Sir Karl Popper, one of the leading philosophers of science of the Twentieth Century, identified intersubjective testability as a cornerstone of science (72). The problem addressed by this requirement is the subjective nature of scientific inference. Because of that subjectivity, conclusions from any individual study may be impacted by unique errors of measurement and bias, which can only be mitigated by a demonstration that the same results hold when the hypothesis is tested under different conditions and by other investigators. Bradford Hill (15) termed this the requirement for "consistency:"

> The lesson here is that broadly the same answer has been reached in quite a wide variety of situations and techniques. In other words, we can justifiably infer that the association is not due to some constant error or fallacy that permeates every inquiry.

Popper's standard for intersubjective testability is met by replication of study results—that is, the same results hold when the hypothesis is tested under the same conditions—and reproducibility—that is, the same results hold when tested under different conditions (73). In forensic pattern comparison disciplines, the claim that a method is a valid means to make source identifications requires testing by multiple researchers/laboratories using a variety of testing paradigms to overcome subjective errors and biases. That consensus can only come from an accumulation of research studies across which stimuli, behavioral procedures, analyses, and investigators are varied, but the results all provide support for the underlying hypothesis.

In the area of firearms and toolmarks, the intersubjective testability guideline has not been met in two critical ways. First, researchers have generally failed to provide the level of detail needed for independent examiners to replicate their studies. Second, and more problematically, the researchers almost exclusively belong to the guild of FATM examiners. We consider each of these issues in turn.

A fundamental tenet of reporting research results is that other researchers could repeat the study in full detail. Unfortunately, this has not happened in much of the firearms and tool mark literature. A striking example is a study by Smith et al. (74), who sent examiners a packet of bullets and cartridge cases along with a blank response sheet and instructions to "evaluate the evidence in its entirety and report any and all conclusions reached on the how the evidence compares to other like items within the sample set, i.e., include all ID's, eliminations, and inconclusives as appropriate." The researchers attempted to decipher how many comparisons were actually made based on what participants wrote on the response sheet; however, this approach led to a number of comparisons that were not possible, and the researchers resorted to speculating about what comparisons the examiners made in the study. Thus, not only could other researchers not replicate the Smith et al. study, but it is unlikely Smith et al. could replicate their own study. Similarly, a study done by Keisler et al. (33) contained numerous ambiguities regarding how the testing was carried out, thus not permitting other researchers to replicate their findings. For instance, the researchers did not explain what pretesting was done, thus not providing information about the difficulty of the tasks. Moreover, the researchers did not control whether subjects completed the test employing their respective laboratory policies or whether they completed the assignment alone or received feedback or assistance from colleagues.

The second way that the firearms and toolmark literature has failed intersubjective testability is the lack of testing done by those without a stake in the outcome of the research. In medical research on a drug's efficacy, for example, relying solely on research conducted by the drug's manufacturer would be considered inherently problematic and has been shown to lead to proindustry results more frequently than when noninterested scientists conducted the study (75, 76). The same should be true for research on forensic pattern comparison methods. Much of the research in this area is carried out by forensic examiners with a horse in the race, most of whom are associated with forensic labs affiliated with law enforcement. Most of the research has been published in the AFTE trade journal.

This latter point illustrates an issue that has been misunderstood and mischaracterized since the 2016 PCAST Report (4). PCAST stated that an important consideration in examining a field's foundational validity is "that a method has been subjected to empirical testing by multiple groups, under conditions appropriate to its intended use." Referring specifically to the FATM literature, PCAST stated as follows:

> The scientific criteria for foundational validity require appropriately designed studies by more than one group to ensure reproducibility. Because there has been only a single appropriately designed study, the current evidence falls short of the scientific criteria for foundational validity. There is thus a need for additional, appropriately designed black-box studies to provide estimates of reliability.

This statement has since been misunderstood to suggest that PCAST believed that just one more study was needed to validate the entire field of firearms identification, despite the call for additional "studies." Indeed, Eric Lander, the cochair of PCAST, lent ammunition to this misunderstanding when he wrote in a law review article, "With only a single well-designed study estimating accuracy, PCAST judged that firearms analysis fell just short of the criteria for scientific validity, which requires reproducibility. A second study would solve this problem" (77). This led the scientifically naive to say that once a second study was completed, the work of the field was done (57).

While we might agree that a narrow hypothesis could gain support from two well-conducted independent studies, validating an entire field with two studies—however well-conducted—is not possible. Indeed, this claim is often associated with the one black box study available in 2016, noted by PCAST in the quote above. That study, though "appropriately designed," considered only cartridge-case comparisons. Could a second study on cartridge-case comparisons validate the entire field of firearms and toolmarks? Consider that the field includes comparisons of cartridge

cases of different materials (copper or steel), cartridge cases fired from a vast array of different guns, bullets involving different materials and guns, and so forth. Added to the extraordinary range of materials involved in pattern-matching bullets and cartridge cases are the differences in levels of training and experience among examiners and the variability of protocols among laboratories for carrying out this work. The suggestion that two studies could validate a field as vast as firearms and toolmarks is absurd on its face.

Reasoning from group data to statements about individual cases. A fundamental disconnect between science and law concerns the application of group-level data to an individual case. Science deals in data that reveal general principles of the natural world; the law deals in inferences about a particular case. Although this issue of generalization from recognized scientific principles to individual cases has been recognized and extensively studied in other scientific domains, such as medical causation or risk assessments, it has been largely ignored in the context of forensic feature-comparison methods. Examiners unwittingly reason from general, group-level data to individual cases ("G2i") without recognizing it as such (6, 78).

Perhaps the most salient example of this G2i problem occurs when discussing error rates. For example, a judge may ask "The examiner concluded an identification; what are the chances of a false positive error in the instant case?" Based on studies purporting to find a 1% false positive error rate, a common response is 1%. However, this response is either a logical failure [sometimes referred to as base rate neglect, the "transposition fallacy" or "prosecutor's fallacy" (79)] or it is premised on an unstated and untenable assumption about other information. A classic example used in medical education nicely illustrates this principle (80).

> Suppose a diagnostic test for cancer has a 1% false positive and false negative error rate. Further, suppose that this test is applied to a population with a base rate for cancer of 500/1,000. Now imagine the test comes back positive; what are the chances this person does not actually have cancer?
>
> The test is 99% accurate—so, it will catch 495 of the 500 with cancer. However, it commits an error 1% of the time so that it will come back positive for 5 of the 500 without cancer. Thus, given the positive test, the probability that this person who tested positive for cancer does not actually have cancer (i.e., a false positive) is 5/500 or 1%.
>
> Now imagine that the same test is applied to a population in which 1 in 1,000 individuals have cancer. Again, the test comes back positive. Now, what are the chances that this person does not have cancer? The test will probably catch the one person with cancer, but it will also come back positive for 1% of the 999 (or about 10 people) who do not have cancer. Thus, given the positive test result, the probability that this person who tested positive for cancer does not actually have cancer (i.e., a false positive) is about 10/11 or 91%.

As is apparent, the posterior probability—the probability that the hypothesis (cancer) is true given the evidence (positive test result)—highly depends on the base rate for the specific disease. The test in the second scenario was no less valid than in the first scenario. Yet, the resulting probability of a false positive is wildly discrepant: 1% vs. 91%!

This example illustrates two different principles. First, it demonstrates the transposition fallacy in assuming a test with a 1% error rate means there is a 1% chance that a positive test result is an error. Second, it illustrates the impact of the base rate on the answer to the question about the probability of an error in the instant case. In laboratory studies, the base rates of same-source and different-source comparisons are arbitrarily assigned by researchers and not applicable to casework. Notably, DNA analysts are generally not permitted to report posterior probability estimates because the base rate is indeterminate and because a base rate or prior probability is beyond the scope of science (81).

The indeterminacy of the base rates suggests that FATM examiners ought to be limited to making general group-level statements, not individualized statements. An example of such testimony might be "the bullet that killed the victim is consistent with having been shot from a 38 caliber Smith & Wesson, and there are approximately 10,000 such guns in circulation in the Southwest United States. Any one of those 10,000 guns could have left similar striae found on the bullet." Provided that the expert has a factual basis for this estimate, one should note that this is still powerfully probative evidence: It reduces the universe of possible guns from over one billion firearms on Earth to a class of 10,000 (82). Future research on the frequency of different striae might enable examiners to reduce further the class of 10,000 Smith & Wesson guns in the Southwest United States. As it stands today, however, this research is a long way off. As a result, FATM examiners cannot provide even basic estimates, such as the number of Smith & Wesson guns in circulation, let alone in a particular jurisdiction at any given time.

Conclusion

Federal Rules of Evidence 702 is set to be amended in December 2023 (83). This amendment—the first in 20 years—was occasioned due to concerns about judges failing to fulfill their gatekeeping function. The amended Rule 702 clarifies the evidentiary standard that must be met (i.e., a preponderance of the evidence) by the proponent of the evidence and emphasizes that judges are responsible for assessing the admissibility of expert testimony rather than permitting the testimony and allowing the jury to determine its weight. This amendment ensures that judges will increasingly be forced to confront the quality of scientific evidence. The guidelines described in this article are primarily intended to establish a scientific baseline for the study of forensic feature-comparison methods. This baseline should help inform how research is conducted in this area of scientific inquiry. Ultimately, however, judges are obligated to evaluate whether expert testimony is based on good grounds. These guidelines, therefore, should serve the secondary purpose of directing judges' inquiries as they fulfill their gatekeeping responsibilities.

Data, Materials, and Software Availability. There are no data underlying this work.

1. M. J. Saks, J. J. Koehler, The coming paradigm shift in forensic identification science. *Science* **309**, 892–895 (2005).
2. B. L. Garrett, N. Scurich, W. E. Crozier, Mock jurors' evaluation of firearm examiner testimony. *Law Hum. Behav.* **44**, 412–428 (2020).
3. National Research Council, *Strengthening Forensic Science in the United States: A Path Forward* (National Academies Press, Washington, DC, 2009).
4. President's Council of Advisors on Science and Technology, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.
5. M. J. Saks, J. J. Koehler, The individualization fallacy in forensic science evidence. *Vand. L. Rev.* **61**, 199–224 (2008).
6. J. Monahan, L. Walker, Social science research in law: A new paradigm. *Am. Psychol.* **43**, 465–498 (1988).
7. *Daubert v., Merrell Dow Pharmaceuticals Inc.*, 509 U.S. 579 (1993).
8. H. J. Rakoff, Keynote address: The future of crime labs and forensic science. *Houston L. Rev.* **57**, 475–481 (2020).
9. D. Kennedy, Forensic science: Oxymoron? *Science* **302**, 1625–1625 (2003).
10. B. L. Garrett, E. Tucker, N. Scurich, Judging firearms evidence. *South. Calif. Law Rev.* **97** (forthcoming).
11. K. M. Lesciotto, The impact of Daubert on the admissibility of forensic anthropology expert testimony. *J. Forensic Sci.* **60**, 549–555 (2015).
12. M. J. Saks, The legal and scientific evaluation of forensic science (especially fingerprint expert testimony). *Seton Hall. L. Rev.* **33**, 1167 (2002).
13. S. I. Gatowski *et al.*, Asking the gatekeepers: A national survey of judges on judging expert evidence in a post-Daubert world. *Law Hum. Behav.* **25**, 433–458 (2001).
14. D. L. Faigman, N. Scurich, T. D. Albright, The field of firearms identification is flawed. *Sci. Am.* https://www.scientificamerican.com/article/the-field-of-firearms-forensics-is-flawed/ (2022).
15. S. B. Hill, Environment and disease: Association or causation? *Proc. R. Soc. Med.* **58**, 295–300 (1965).
16. K. M. Fedak, A. Bernal, Z. A. Capshaw, S. Gross, Applying the Bradford Hill criteria in the 21st century: How data integration has changed causal inference in molecular epidemiology. *Emerg. Themes Epidemiol.* **12**, 1–9 (2015).
17. R. R. Neutra, C. F. Cranor, D. Gee, The use and misuse of Bradford Hill in US Tort Law. *Jurimetrics* **58**, 127–162 (2018).
18. J. Sanders, D. L. Faigman, P. B. Imrey, P. Dawid, Differential etiology: Inferring specific causation in the law from group data in science. *Arizona Law Rev.* **63**, 851–922 (2021).
19. B. McClintock, The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U.S.A.* **36**, 344–355 (1950).
20. S. Ravindran, Barbara McClintock and the discovery of jumping genes. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 20198–20199 (2012).
21. *Glossary of the Association of Firearm and Tool Mark Examiners* (ed. 6) (2013). https://afte.org/uploads/documents/AFTE_Glossary_Version_6.091922_FINAL_COPYRIGHT.pdf.
22. P. H. Lindsay, D. A. Norman, *Human Information Processing: An Introduction to Psychology* (Academic Press, New York, 1977).
23. I. Biederman, M. M. Shiffrar, Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *J. Exp. Psychol. Learn. Mem. Cogn.* **13**, 640–645 (1987).
24. J. L. Mnookin, Of black boxes, instruments, and experts: Testing the validity of forensic science. *Episteme* **5**, 343–358 (2008).
25. T. D. Cook, D. T. Campbell, W. Shadish, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton Mifflin, Boston, MA, 2002).
26. S. S. Diamond, "Reference guide on survey research" in *Reference Manual on Scientific Evidence* (The National Academies Press, Washington DC, ed. 3, 2000).
27. S. Messick, Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* **50**, 741–749 (1995).
28. D. T. Campbell, Factors relevant to the validity of experiments in social settings. *Psychol. Bull.* **54**, 297–312 (1957).
29. A. Schwartz, A systemic challenge to the reliability and admissibility of firearms and toolmark identification. *Colum. Sci. Techol. Law Rev.* **6**, 1–64 (2004).
30. J. E. Hamby, D. J. Brundage, N. D. Petraco, J. W. Thorpe, A worldwide study of bullets fired from 10 consecutively rifled 9 MM RUGER pistol barrels–analysis of examiner error rate. *J. Forensic Sci.* **64**, 551–557 (2019).
31. R. G. Nichols, Defending the scientific foundations of the firearms and tool mark identification discipline: Responding to recent challenges. *J. Forensic Sci.* **52**, 586–594 (2007).
32. D. P. Baldwin, S. J. Bajic, M. Morris, D. Zamzow, "A study of false-positive and false-negative error rates in cartridge case comparisons" (US Department of Energy, Ames Laboratory, Iowa, 2014), https://apps.dtic.mil/sti/pdfs/ADA611807.pdf.
33. M. A. Keisler, S. Hartman, A. Kilmon, M. Oberg, M. Templeton, Isolated pairs research study. *AFTE J.* **50**, 56–58 (2018).
34. K. L. Monson, E. D. Smith, E. M. Peters, Accuracy of comparison decisions by forensic firearms examiners. *J. Forensic Sci.* **68**, 86–100 (2023).
35. B. A. Best, E. A. Gardner, An assessment of the foundational validity of firearms identification using ten consecutively button-rifled barrels. *AFTE J.* **54**, 28–37 (2022).
36. M. Guyll, S. Madon, Y. Yang, K. A. Burd, G. Wells, Validity of forensic cartridge-case comparisons. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2210428120 (2023).
37. N. Scurich, B. L. Garrett, R. M. Thompson, Surveying practicing firearm examiners. *Forensic Sci. Int. Synerg.* **4**, 100228 (2022).
38. M. Sinha, R. E. Gutierrez, Signal detection theory fails to account for real-world consequences of inconclusive decisions. *Law Prob. Risk* **21**, 131–135 (2022).
39. I. E. Dror, N. Scurich, Mis)use of scientific measurements in forensic science. *Forensic Sci. Int. Synerg.* **2**, 333–338 (2020).
40. T. J. Weller, M. D. Morris, Commentary on: I Dror, N Scurich, "(Mis) use of scientific measurements in forensic science". *Forensic Sci. Int. Synerg.* **2**, 701 (2020).
41. N. Scurich, I. E. Dror, Continued confusion about inconclusives and error rates: Reply to Weller and Morris. *Forensic Sci. Int. Synerg.* **2**, 703–704 (2020).
42. H. R. Arkes, J. J. Koehler, Inconclusives are not errors: A rejoinder to Dror. *Law Prob. Risk* **21**, 89–90 (2022).
43. N. Scurich, Inconclusives in firearm error rate studies are not "a pass". *Law Prob. Risk* **21**, 123–126 (2022).
44. A. H. Dorfman, R. Valliant, Inconclusives, errors, and error rates in forensic firearms analysis: Three statistical perspectives. *Forensic Sci. Int. Synerg.* **5**, 100273 (2022).
45. S. A. Cole, B. C. Scheck, Fingerprints and miscarriages of justice: Other types of error and a post-conviction right to database searching. *Alb. Law Rev.* **81**, 807–850 (2017).
46. N. Scurich, R. S. John, Three-way ROCs for forensic decision making. *Stats. Pub Pol'y* 1–16 (2023).
47. G. T. Chiodo, S. W. Tolle, L. Bevan, Placebo-controlled trials: Good science or medical neglect? *Western J. Med.* **172**, 271–273 (2000).
48. B. L. Garrett, G. Mitchell, The proficiency of experts. *Univ. Pennsylvania Law Rev.* **166**, 901–948 (2017).
49. B. Max, J. Cavise, R. E. Gutierrez, Assessing latent print proficiency tests: Lofty aims, straightforward samples, and the implications of nonexpert performance. *J. Forensic Identif.* **69**, 281–298 (2019).
50. R. Balko, Devil in the grooves: The case against forensic firearms analysis. Available at: https://radleybalko.substack.com/p/devil-in-the-grooves-the-case-against.
51. J. M. Tangen, M. B. Thompson, D. J. McCarthy, Identifying fingerprint expertise. *Psychol. Sci.* **22**, 995–997 (2011).
52. C. Bird, B. Found, D. Rogers, Forensic document examiners' skill in distinguishing between natural and disguised handwriting behaviors. *J. Forensic Sci.* **55**, 1291–1295 (2010).
53. D. G. Mook, In defense of external invalidity. *Am. Psychol.* **38**, 379–387 (1983).
54. T. D. Albright, Perceiving. *Daedalus* **144**, 22–41 (2015).
55. T. D. Albright, On the perception of probable things: Neural substrates of associative memory, imagery, and perception. *Neuron* **74**, 227–245 (2012).
56. T. D. Albright, W. A. Freiwald, "High-level visual processing: From vision to cognition" in *Principles of Neural Science*, E. R. Kandel, J. D. Koester, S. H. Mack, S. A. Siegelbaum, Eds. (McGraw-Hill, New York, ed. 5, 2021), pp. 564–581.
57. J. Agar, The admissibility of firearms and toolmarks expert testimony in the shadow of PCAST. *Baylor Law Rev.* **93**, 196–220 (2022).
58. K. N. Ballantyne, G. Edmond, B. Found, Peer review in forensic science. *Forensic Sci. Int.* **277**, 66–76 (2017).
59. W. Thompson, J. Black, A. Jain, J. Kadane, *Forensic Science Assessments: A Quality and Gap Analysis–Latent Fingerprint Examination* (American Association for the Advancement of Science, Washington, DC, 2017).
60. W. C. Thompson, Decision thresholds, contextual bias and the accuracy of verdicts. *Proc. Natl. Acad. Sci. U.S.A.* this issue, 2023–01844 (2023).
61. M. L. Pierce, L. J. Cook, Development and implementation of an effective blind proficiency testing program. *J. Forensic Sci.* **65**, 809–814 (2020).
62. J. J. Koehler, Proficiency tests to estimate error rates in the forensic sciences. *Law Prob. Risk* **12**, 89–98 (2013).
63. C. Hundl, M. Neuman, A. Rairden, P. Rearden, P. Stout, Implementation of a blind quality control program in a forensic laboratory. *J. Forensic Sci.* **65**, 815–822 (2020).
64. R. Mejia, M. Cuellar, J. Salyards, Implementing blind proficiency testing in forensic laboratories: Motivation, obstacles, and recommendations. *Forensic Sci. Int. Synerg.* **2**, 293–298 (2020).
65. I. Etikan, K. Bala, Sampling and sampling methods. *Biometr. Biostat. Int. J.* **5**, 00149 (2017).
66. H. H. Dodge, Y. Katsumata, J. A. Kaye, Characteristics associated with willingness to participate in a randomized controlled behavioral clinical trial using home-based personal computers and a webcam. *Trials* **15**, 1–7 (2014).
67. S. J. Bajic, L. S. Chumbley, M. Morris, D. Zamzow, "Report: Validation study of accuracy, repeatability, and reproducibility of firearms comparisons" (Ames Laboratory-USDOE Technical Report # ISTR-5220, 2020).
68. A. Karni, D. Sagi, The time course of learning a visual skill. *Nature* **365**, 250–252 (1993).
69. D. Nunan, J. Aronson, C. Bankhead, Catalogue of bias: Attrition bias. *BMJ Evid. Based Med.* **23**, 21–22 (2018).
70. J. C. Dumville, D. J. Torgerson, C. E. Hewitt, Reporting attrition in randomised controlled trials. *British Med. J.* **332**, 969–971 (2006).
71. J. J. Koehler, S. Liu, Fingerprint error rate on close non-matches. *J. Forensic Sci.* **66**, 129–134 (2021).
72. K. Popper, *The Logic of Scientific Discovery* (Routledge, 2005).
73. J. Freese, D. Peterson, Replication in social science. *Ann. Rev. Soc.* **43**, 147–165 (2017).
74. T. P. Smith, A. Smith, J. B. Snipes, A validation study of bullet and cartridge case comparisons using samples representative of actual casework. *J. Forensic Sci.* **61**, 939–946 (2016).
75. J. E. Bekelman, Y. Li, C. P. Gross, Scope and impact of financial conflicts of interest in biomedical research: A systematic review. *JAMA* **289**, 454–465 (2003).
76. J. P. Singh, M. Grann, S. Fazel, Authorship bias in violence risk assessment? A systematic review and meta-analysis. *Plos One* **9**, e72484 (2013).
77. E. S. Lander, Fixing Rule 702: The PCAST Report and steps to ensure the reliability of forensic feature-comparison methods in the criminal courts. *Fordham L. Rev.* **86**, 1672–1692 (2018).
78. D. L. Faigman, J. Monahan, C. Slobogin, Group to individual (G2i) inference in scientific expert testimony. *Univ. Chicago Law Rev.* **4**, 417–480 (2014).
79. W. C. Thompson, E. L. Schumann, Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law Hum. Behav.* **11**, 167–187 (1987).
80. N. Scurich, R. S. John, Prescriptive approaches to communicating the risk of violence in actuarial risk assessment. *Psychol. Pub. Pol'y L.* **18**, 50–78 (2012).
81. W. C. Thompson, J. Vuille, A. Biedermann, F. Taroni, The role of prior probability in forensic assessments. *Front. Genet.* **4**, 1–3 (2013).
82. https://www.smallarmssurvey.org/database/global-firearms-holdings.
83. https://www.uscourts.gov/rules-policies/archives/committee-reports/advisory-committee-evidence-rules-may-2022.