JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, July 2020, Vol. 65,
doi: 10.1111/1556-4029.14435
Available online at: onlinelibrary.wiley.com

COMMENTARY

Itiel E. Dror,[1,2] *Ph.D.*

The Error in "Error Rate": Why Error Rates Are So Needed, Yet So Elusive

ABSTRACT: Establishing error rates is crucial for knowing how well one is performing, determining whether improvement is needed, measuring whether interventions are effective, as well as for providing transparency. However, the flurry of activities in establishing error rates for the forensic sciences has largely overlooked some fundamental issues that make error rates a problematic construct and limit the ability to obtain a meaningful error rate. These include knowing the ground truth, establishing appropriate databases, determining what counts as an error, characterizing what is an acceptable error rate, ecological validity, and transparency within the adversarial legal system. Without addressing these practical and theoretical challenges, the very notion of a meaningful error rate is limited.

KEYWORDS: forensic science, error rates, inconclusive, validation, proficiency testing, Daubert, quality assurance.

Every domain should have error rates, because it is one of the fundamental metrics that allows to quantify performance.

Why Error Rates Are So Needed

Having established error rates enables to:

Know How Well One is Performing

Without error rates, one has limited insights to how well they are performing in terms of errors. They may feel they are doing great, but their actual performance may be poor, or conversely, they may feel they are performing poorly, but actually they may be performing well. Generally, reflecting introspectively about one's own performance, "meta-cognition," is problematic and weak (e.g., [1]). For example, the correlation between confidence and accuracy is low (e.g., [2]). Hence, people can feel very confident, but nevertheless be erroneous, and the reverse, feel very unconfident but nevertheless be correct (3).

Such "blind spots" in estimating one's own performance are prevalent with humans (4). Forensic examiners are no exception; they are not immune to the vulnerability of blind spots in evaluating their own performance (5), including when it comes to perception and estimations of errors and error rates (6).

Hence, reliance on internal metrics is generally limited and can be misleading, and in science it is unacceptable. Proper external metrics must be used to determine and measure error rate and performance, but these—at least in science—need to be observable, quantifiable, and replicable. Furthermore, even external metrics can be fundamentally flawed, such as being content with performance just because the "court accepts it", which is far from being scientifically appropriate, nor is it acceptable from an epistemological perspective.

Determine Whether Improvement is Needed

In order to decide whether any remedial action, intervention, or changes are needed, one must examine current error performance against optimal performance or a target performance level of errors. Without metrics, one cannot determine whether improvement is needed, or even possible. Unbeknown optimal error rates (or the best ones practically achievable) might have already been reached, and any further attempted improvement will be futile, a waste of time and money, and just bring about unnecessary frustration.

Measure Whether Interventions are Effective

When measurements are taken to improve error performance, error rates allow to examine, assess, and quantify the effectiveness of these interventions. Without measurements, one cannot know whether errors have decreased, remained unchanged, or increased (see above, for requiring proper error rates, and not, for example, relying on confidence, introspection, or the courts).

Provide Transparency to Users or Consumers

It is hard for a consumer to know how to assess, rely on, and use the product when an error rate is not known. In the forensic domain, there are many users and stakeholders (e.g., investigative, judicial, regulatory and public—See figure 1 in [7]). For example, without error rates how can the court know how to consider and weigh the evidence presented (8,9)? Jurors often rely on their knowledge of forensic science via entertainment programs on television, such as "CSI" (10,11), or based on the confidence in which the expert presents their evidence (see above, as to the flaw of such reliance). Established error rates are a vital tool for the court, the investigators, and other fact

[1]UCL Centre for the Forensic Sciences, University College London, London, U.K.
[2]Cognitive Consultants Internationals (CCI-HQ), London, U.K.
Corresponding author: Itiel Dror, Ph.D. E-mail: i.dror@ucl.ac.uk

Received 16 Feb. 2020; and in revised form 22 Mar. 2020; accepted 24 Mar. 2020.

finders and users. Without this information, they may be unintentionally misled.

These are all important reasons to establish error rates. Indeed, the National Academy of Sciences report on forensic science (12), as well as the President's Council of Advisors on Science and Technology report on forensic science (13), the National Commission on Forensic Science (14), and the National Institute of Standards and Technology Expert Group report (15), have all called for the forensic science domains to establish their error rates, as these were largely unknown.

Since then there has been a flurry of activity in attempts to establish error rates, from forensic domains, such as firearms (e.g., [16-20]), fingerprinting (e.g., [21,22]), anthropology (e.g., [23,24]), and DNA (e.g., [25-27]), to domains such as duct tape (e.g., [28]), forensic entomology (e.g., [29]), blood pattern analysis (e.g., [30]), bitemarks (e.g., [31-33]), ear marks (e.g., [34]), and even printing defects (e.g., [35]); as well as error perception (e.g., [6])—all important endeavors. However, as can be seen from these references, most of these efforts have taken place within the last 10–15 years, predominantly since the publication of the NAS report in 2009 (12).

However, establishing error rates in forensic science is not as straightforward as it may seem. There are a myriad of problems and limitations, some of them specific to forensic science that includes practical and theoretical challenges.

Why Are Error Rates So Elusive

Knowing the Ground Truth

To establish error rates, one must know the ground truth as a matter of fact. Expert consensus and past cases, where the courts have established guilt "beyond reasonable doubt," are not appropriate for determining error rates. Error rates must be based on testing stimuli for which the ground truth is known and should be cataloged and maintained in databases. This is paramount, but hardly existed in the past, although they are now starting to emerge. These efforts require, for example in fingerprinting, asking volunteers to submit their fingerprints as reference data, having them touch objects and surfaces, and then recording the latent prints they deposit so a database can be generated with the ground truth of which test pairs (a questioned and a known mark) are indeed a match as a matter of fact.

However, creating such ground truth databases is more difficult in other forensic domains. Consider, for example, a bloodstain pattern database: It is not feasible to take human subjects and injure them so as to create ground truth bloodstain patterns from various impact events—for example, shoot them in order to create a ground truth gunshot blood spatter. Such limitations can perhaps be partially overcome by using real cases where the ground truth is known, or computer simulations. However, each of these has limitations, which need to be specified when communicating the error rate.

Establishing Appropriate Databases

There are numerous theoretical and practical issues with establishing ground truth databases.

First, the items to be included in the database must also be examined and approved by human expert examiners—not to determine the ground truth (which is known), but to determine what decision conclusion can be justifiably reached. This is because the mere fact that we know with certainty how a particular sample was manufactured does not mean that the generated test sample actually allows for an examiner to arrive at the ground truth, in which case an inconclusive decision is the correct conclusion. This could be, for example, due to the quality of the sample being insufficient or having ambiguity in the questioned sample. Therefore, samples selected for testing examiner performance must undergo a testing and approval process to establish the appropriate and correct decision. This, of course, raises challenging questions about who and how, and what criteria, will be used to determine the correct conclusion.

Second, ground truth databases should include appropriate "nonmatches," as the databases for establishing error rates need to have both matches and nonmatches. The nonmatches need not only be ground truth nonmatches, but also need to include challenging and difficult cases, "look alikes" that are nevertheless a nonmatch. If, for example, in a fingerprint test all nonmatches involve samples where the fingerprints in question are very different and distinct from the known prints (i.e., there are obvious discrepancies), then the resulting error rate will be artificially reduced, as there are no challenging cases in the database. With firearms and fingerprinting, this can be achieved by obtaining similar nonmatches though computerized software of existing individual characteristic databases, such as the National Integrated Ballistic Information Network (NIBIN) or Automated Fingerprint Identification Systems (AFIS), which examine a huge number of marks and provide look alike similar patterns. However, getting such challenging nonmatches is considerably more difficult in other forensic domains.

Third, there is an issue as to what level of difficulties should be included in the database. This does not only pertain to nonmatches (see above), but even the matches can be easier or more difficult. For example, a ground truth firearms evidence match that has high quality and quantity of information will be an easier match to declare than a firearms evidence with less information, distortions, noise, and other artifacts. Furthermore, the pair of patterned items (e.g., the known breech face pattern created from the test fire in the laboratory and the unknown breech face pattern from the crime scene) can include, or not, varying distinctive and informative features, making it easier or more difficult to declare a match.

Levels of difficulty in making a determination influence error rates, and the distribution of difficulties needs to represent correctly that which exists in real casework. If the database includes matches and nonmatches that are too easy, then the error rates established will be artificially and incorrectly low (see above). However, if the database includes too many difficult and challenging cases, then the error rates established will be artificially and incorrectly high. The databases and any testing (e.g., proficiency testing) need to be a representative sample of the population of real casework (36). This means that one must first determine the distribution of difficulties in the real world of forensic work, and the database must mimic and be representative of those difficulties. This is not only a task that requires serious effort, but also has theoretical challenges, such as how to quantify difficulty (i.e., prior to starting practical work in constructing the databases, assessing the difficulty in real casework and then assessing the level of difficulty for each test items to be included in the database).

The correct distribution of cases does not only pertain to the level of difficulty (i.e., how hard it is to make the decision, for both a "match" and a "nonmatch"), but also to the source of the difficulty (e.g., quantity and quality of information, distinctiveness, noise, distortions, etc.).

Fourth, there is the question of how to distribute other parameters in the database beyond the level of difficulty (see above). For example, in fingerprinting, shall the database sets contain latent prints equally from all 10 fingers, or more from the fingers that are actually used more and are likely to be found in real cases, that is, the Pollex, and the Digiti Me'dius and Secundus Manus (Thumb, Middle and Index fingers) and less of the Digitus Mi'nimus Ma'nus (the pinky)?

On the one hand, the database should reflect, be a sample of, the fingerprints in real case work, and hence should not have an equal number of latent prints from each finger (and, similarly, in firearms, for example, the database should include breech faces from calibers and guns as a representative sample of the populations of those in casework).

But, on the other hand, how far shall the databases go with this approach? For example, with fingerprints, shall it contain many more prints from the right hand (given that most people are right handed, so only about 10% of the database samples will be of fingers from the left hand)? And, to push this further, given that most cases in crime involve men, shall the databases mostly include men? Shall the database also reflect the distribution from casework of different age groups? Races? Etc. --Not a simple matter.

The issues raised in this paper apply in one way or the other to all forensic domains. For example, DNA cases also involve many factors to consider. Not only are there simple cases (i.e., single source samples for comparison) and difficult cases (e.g., DNA mixtures), but there are increasing levels of difficulty even within mixture DNA cases. For example, a four-person mixture is very different than a two-person mixture, and even a two-person mixture in equal proportions from both contributors is different than a two-person mixture that has a major contributor and a minor contributor; there can also be allelic dropout, stutter, etc.

The multiple factors involved in DNA, as well as other forensic domains, each adds complexity and difficulty to the analysis, which impacts the error rate, and requires that we overcome theoretical and practical issues in constructing databases to determine error rates. Theoretically, one must figure out all the factors that characterize cases and their difficulty, and how these can be measured and quantified. Then, one needs to practically determine the distribution of these factors in real casework, so as to construct a database that mimics the real world (and also to determine which factors should mimic the real world—age, race, sex, etc.—see above). Complex issues underpin the very notion of error rates.

Fifth, another monumental challenge in creating ground truth databases for establishing error rates is that even if we know the ground truth, as we created the database, it does not mean that there is only one possible cause or source. For example, for a bloodstain pattern database, we might aim to create a bloodstain pattern using a gun shot from a specific angle. However, it may be impossible to know whether that same blood pattern could not also be caused by a different angle, or maybe even a different weapon altogether. Hence, even if we know the ground truth, there may be more than one set of circumstances that could create this pattern, and thus more than one possible answer. Similarly, different fingers can create the same latent fingermark (that is not to say that fingerprints are not unique, but latent fingermarks are only a partial impression of the full fingerprint; hence, fingerprints may have certain segments that are the same, and can therefore create the same latent mark).

If there is more than one possible cause for the pattern, (which is very hard to rule out and exclude), then this raises the issue of how to determine what is a correct decision versus what is an error. If during testing for determining an error rate, the examiner reaches the conclusion of what we know to be the ground truth, it may nevertheless still be an error as the same pattern can be caused by a different source. Conversely, if the examiner reaches a conclusion that differs from the ground truth, how can we count that as an error, if the same pattern can also be correctly attributed to a different source. If and when we know of such situations, then perhaps the only correct answer is inconclusive—but then we count a ground truth right answer as an error, because there are other possible right answers, making inconclusive the only appropriate answer. The major challenge to accurately estimating an error rate is that we may not be aware of the existence of such situations.

Furthermore, the option to try and address this issue by characterizing such cases as "inconclusives" is problematic in a number of other ways: how can we ever determine that no other source or event can produce the same pattern or mark? Even if we do find a way to determine that, then do we simply not include inconclusives in the databases? And, if we include inconclusives, what do we do with them in terms of counting an error rate? This leads to the next issue of what counts as an error.

Determining What Counts as an Error

Very simply put, an error is when an incorrect decision is made given the information available. However, there are many complex issues to address in determining what counts as an error and the use of the term error (37), which underpins establishing error rates. To start, are all errors equal? Is a false positive (an erroneous identification) not a more serious error than a false inconclusive (a decision that the examination is inconclusive, cannot decide, while there is clear information to make a decision)? That is not a simple issue and must be addressed for determining error rates. Clearly, deciding not to decide is a decision—there are cases where such a decision is correct, and there are cases when such a decision is erroneous (38).

One can claim that erroneous inconclusives should not to be included in calculating error rates, as error rates will be used in court to assess identification decisions, and therefore, error rates should only be based on false positives. Furthermore, the consequences of a false identification (potentially sending an innocent person to jail) are considered by many as the most serious error, and therefore is the type of error to be used in calculating error rates. However, there are also potentially serious consequences for a false inconclusive (letting a guilty person go free, or not excluding an innocent person). Furthermore, if flawed inconclusive decisions are not considered as errors, or of a lesser error, then examiners can artificially and incorrectly reduce the error rate estimates by resorting to inconclusive decisions more often than they do in real cases.

A potential solution is to avoid this quandary by not including inconclusive decisions in the databases. However, this undermines the databases as reflecting the real world of forensic casework, as well as the ecological validly of error rates (see below).

To make the issue of determining what counts as an error even more complex, even though the ground truth is known in the databases for determining error rates, that does not mean that an identification decision is correct even when it is a "match": The information available may make an inconclusive decision the correct decision, and a match decision incorrect, even if the questioned and the known samples do actually come from the

same source (38). Recall our earlier example, that two unique fingerprints can deposit the same partial latent mark, hence the mere fact that we know with certainty how a particular sample was manufactured, does not mean that the deposited latent mark allows for an examiner to justifiably arrive at the ground truth conclusion of a match.

Characterizing What is an Acceptable Error Rate

Even if we overcome all these difficulties and establish an error rate, then what is an acceptable error rate? Zero error rates, claimed in the past in forensic science (e.g., [39]) are not only unrealistic and a misrepresentation of the science, but can cause harm to individual examiners and to the field as a whole (this is not only true in forensic science, but in other domains, see, e.g., [40]).

We can use an error rate measure to assess progress and improvements, but what is an acceptable error rate is not a simple matter. There are inconsistencies across industries. For instance, in some industries, such as healthcare, higher error rates—even resulting in death—are more acceptable relative to other industries, such as aviation, where even the lowest errors resulting in death are deemed unacceptable. Even within industries, there can be inconsistency; for example, one laboratory may be content with a certain error rate while another laboratory may feel that same error rate is unacceptable. In forensic science, there are no standards as to what constitutes an acceptable error rate.

One can try to avoid this issue, by letting the courts decide how to conceptualize and deal with error rates. However, if we undertake the monumental task of establishing error rates, then we need to make it useful. How do we expect the jurors to understand and use an error rate? Providing a number, X, as the error rate, needs to be accompanied by guidelines regarding what it means. If one cannot provide such characterization of the error rate number, then perhaps the characterization will be relative to error rates of other forensic laboratories, other examiners, etc. However, this is limited too, as there are different reasons that error rates may vary.

Providing "an error rate" for a forensic domain may be misleading because it is a function of numerous parameters and depends on a variety of factors. An error rate varies by difficulty of the decision (see the discussion about DNA difficulties, above). Error rates are going to be higher for difficult cases, but lower for easier cases. So, a single "error rate" is misleading, and therefore, it may be best to provide error rates as a function of difficulty, which will be helpful, if there is a measure of difficulty for the case at hand, so the correct error rate can be attributed.

An error rate will also vary across individuals. Some experts have higher error rates, and others, lower error rates. This can be a function of training background (not only the quantity and quality of the training, per se, but also different schools of thought and strategies, some more conservative than others, etc.), as well as cognitive aptitude, motivation, ideology, experience, etc. Therefore, error rates may give insights into forensic domains in general, however, may say very little about a specific examiner's decision in a particular case (41).

Hence, an average error rate for an average expert, in an average case, may not be informative (may even be misleading) for evaluating a specific expert examiner, doing a specific case. However, providing ranges, confidence intervals, standard deviations, and other information about "the error rate," as well as comparative error rates of others, may all be helpful in understanding and evaluating a general error rate. Nevertheless, what is an acceptable error rate is a problematic issue, not only can it differ across examiners and case difficulty, but also across forensic laboratories, as well as across courts and jurisdictions (what one court deems as acceptable, may not be acceptable by another court).

Ecological Validity

Ecological validity is critical for an error rate being accurate, meaningful, and helpful (or, error rates, if we have them for different difficulties, different types of decisions, etc.). In addition to having the correct distribution of types of samples in the databases, etc. (discussed earlier), there are a number of other critical issues as per ecological validity. If we want the error rates to reflect real casework, then their establishment must mirror real casework conditions. This means, first and foremost, that examiners should not know they are being tested, and must think they are conducting real casework (7,42,43). It has been well established that when people know they are being tested, and even more so, measuring errors in their decisions, their performance is going to be different (e.g., [44,45]; see also the Hawthorne effect).

Similarly, all other conditions need to be as they are in real case work, from knowing contextual irrelevant information (46), to workload and stress (47). Of course, this poses not only practical challenges, but also theoretical issues, because examiners work on a variety of different cases and circumstances, across different forensic laboratories.

Another issue to consider is that in almost all forensic domains, decisions are verified or reviewed one way or the other. Therefore, if an examiner makes an error it can be corrected prior to the case report being finalized. An ecological valid error rate therefore needs to go through these processes as well. However, here too forensic laboratories vary widely in what decisions are verified (some only verify identifications, whereas others verify all decisions) and how they verify (e.g., some do blind verifications, whereas others do not).

Regardless of how verification is carried out and under what circumstances, of particular interest is how often examiners disagree when they examine the exact same case. This pertains to between (inter) examiner differences when different examiners reach different conclusions, and also to within (intra) examiner differences whereby the same examiner examines the same case at two different times (48). For example, in fingerprinting, 10% of the time the same examiner will reach different conclusions when examining the same pair of fingerprints (21).

Data should be collected routinely on disagreements between examiners. Unfortunately, these opportunities to collect valuable data are missed when disagreements are being resolved without proper documentation. These data are important for establishing error rates, because if verifiers catch errors, then this needs to be accounted for in calculating error rates (i.e., not to include errors that will be fixed during verifications). Conversely, if verifiers most always verify the conclusions (even when an error is made), then there is less need to take the verification into account when calculating error rates.

It is important to have forensic examiners involved in conducting error rate studies, but it is also important to remember that they are the object of examination, which leads to a clear conflict of interest and biases. Their involvement is paramount

and necessary—error rate studies cannot be conducted without their involvement—but it raises concerns anytime one measures their own performance, especially about processes that they have been doing for a long time and in which they are heavily invested.

Transparency Within the Adversarial Legal System

As mentioned earlier, error rates are critical metrics for forensic laboratories to assess and improve performance, and to provide transparency. However, these efforts are easily stifled by the adversarial legal system. Forensic laboratories need to really engage in checking their performance, carry out real quality assurance, look for issues and establish error rates, all in an open and transparent manner—efforts that need to be encouraged and commended. In reality, however, in an adversarial legal system, such undertakings will be used in court to undermine and attack the work of the forensic examiner and the forensic laboratory itself.

This is the nature of the adversarial legal system. Interestingly, the requirement for having error rates actually came from the courts in 1993, in *Daubert v. Merrill Dow Pharmaceuticals*, long before the NAS 2009 report (12). In this case, the US Supreme Court ruled that when considering admissibility in court, the potential or known error rate of the scientific technique in question should be established (Daubert [49]). How can we demand, or even expect, laboratories to carry out such work and be transparent about it, if these will be used against them?

In theory, we can consider establishing two different types of error rates: one, a "summative" error rate, and another, an "informative" error rate. The summative error rate will reflect the error rate so as to serve legal purposes. The informative error rate will be established for quality assurance and improvement purposes. The former will focus on whether a discipline is sufficiently established to present evidence in court (e.g., Daubert hearings), as well as to help evaluate the error rate of a specific examiner, in a specific case, deliberated before the court (thus, e.g., taking into account the difficulty of the case, the examiner's experience, etc.). The latter, the informative error rate, will be established for developing interventions and improvement monitoring. And hence this error rate will be calculated differently (e.g., not take into account the errors detected and corrected during verification). Of course, some parameters will be used by both types of error rates, and there are some theoretical and practical issues with the distinction between these two proposed error rates. However, there is a basic problem and tension between the different aspects and purposes of having an error rate (e.g., quality control, transparency, and legal) that need to be addressed.

Summary and Conclusion

I have outlined many of the difficulties, challenges, obstacles, and inherent limitations in determining error rates in forensic science. Nevertheless, in doing so, I do not wish to undermine or repress these gallant efforts. But, to move forward we need to understand the complexity in determining and providing error rates, and the inherent limitations of the very notion of error rates.

The need to properly establish error rates in forensic science is clear. But, given the time and effort it requires, as well as the inherent limitations of the very notion of error rates, is it worth it? And, how does it compare (or complement) other measures of performance (e.g., effective proficiency testing, quality assurance checks such as dip sampling and blind verification, accreditation, and ongoing training and development). Regardless, it is nevertheless a very worthwhile endeavor, because the benefits of establishing error rates go beyond error rates per se.

Establishing (or even the attempt at trying to establish) error rates involves examination of various issues, which, by itself, will benefit forensic science. This is exactly what happened with the issue of bias: As part of the effort to minimize bias, discussions emerged to determine what is ir/relevant contextual information for doing forensic work (e.g., [50]). Determining what is ir/relevant for forensic work was important and benefited forensic science beyond the issue of bias, per se. Similarly, the efforts in establishing error rates will benefit forensic science. For example, considering and examining how different training, experience, and aptitude of examiners, as well as difficulty and types of decisions, impact performance, are not only relevant to error rates. If these are going to be researched as part of the effort to try and establish error rates, then that will surely bring benefit to forensic science.

Acknowledgments

I want to thank Michal Pierce, Gillian Tully, Hal Stern, Lynn Garcia, Joseph Almog, and Alex Biedermann for their valuable comments on an earlier version of this manuscript.

References

1. Oskamp S. Overconfidence in case-study judgments. J Consult Psychol 1965;29:261–5.
2. Ryback D. Confidence and accuracy as a function of experience in judgment making in the absence of systematic feedback. Percept Mot Ski 1967;24(1):331–4.
3. Bothwell RK, Deffenbacher KA, Brigham JC. Correlation of eyewitness accuracy and confidence. J Appl Psychol 1987;72(4):691–5.
4. Pronin E, Lin DY, Ross L. The bias blind spot: perceptions of bias in self versus others. Pers Soc Psychol Bull 2002;28(3):369–81.
5. Kukucka J, Kassin S, Zapf P, Dror IE. Cognitive bias and blindness: a global survey of forensic science examiners. J Appl Res Mem Cogn 2017;6(4):452–9.
6. Murrie DC, Gardner BO, Kelley S, Dror IE. Perceptions and estimates of error rates in forensic science. Forensic Sci Int 2019;302:109887.
7. Dror IE, Pierce ML. ISO standards addressing issues of bias and impartiality in forensic work. J Forensic Sci 2019;https://doi.org/10.1111/1556-4029.14265.
8. Edmond G. The admissibility of forensic science and medicine evidence under the Uniform Evidence Law. Crim Law J 2014;38:136–58.
9. Koehler J. Fingerprint error rates and proficiency tests: what they are and why they matter. Hastings Law J 2008;59(5):1077–100.
10. Smith SM, Stinson V, Patry MW. Fact or fiction? The myth and reality of the CSI effect. Court Rev 2011;100(47):4–7.
11. Cole SA, Dioso-Villa R. CSI and its effects: media, juries, and the burden of proof. N Engl Law Rev 2007;41(3):435–70.
12. National Research Council. Strengthening forensic science in the United States: a path forward. Washington, DC: The National Academies Press, 2009.
13. President's Council of Advisors on Science and Technology (PCAST). Report to the President – Forensic science in criminal courts: ensuring validity of feature-comparison methods. Washington, DC: Office of Science and Technology, 2016.
14. National Commission on Forensic Science. Facilitating research on laboratory performance. Washington, DC: National Commission on Forensic Science, 2016.
15. National Institute of Standards and Technology (NIST). Latent print examination and human factors: improving the practice through a systems approach. Gaithersburg, MD: National Institute of Standards and Technology, 2012.
16. Cazes M, Goudeau J. Validation study results from hi-point consecutively manufactured slides. AFTE J 2013;45:175–7.

17. Mattijssen E, Witteman C, Berger C, Brand N, Stoel RD. Validity and reliability of forensic firearm examiners. Forensic Sci Int 2020;307:110112.
18. Hamby JE, Norris S, Petraco ND. Evaluation of GLOCK 9 mm firing pin aperture shear mark individuality based on 1,632 different pistols by traditional pattern matching and IBIS pattern recognition. J Forensic Sci 2016;61(1):170–6.
19. Hamby JE, Brundage DJ, Petraco D, Thorpe JW. A worldwide study of bullets fired from 10 consecutively rifled 9MM RUGER pistol barrels — analysis of examiner error rate. J Forensic Sci 2019;64(2):551–7.
20. Smith TP, Smith AS, Snipes JB. A validation study of bullet and cartridge case comparisons using samples representative of actual casework. J Forensic Sci 2016;61(4):939–46.
21. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Repeatability and reproducibility of decisions by latent fingerprint examiners. PLoS ONE 2012;7(3):e32800.
22. Tangen JM, Thompson MB, McCarthy DJ. Identifying fingerprint expertise. Psychol Sci 2011;22(8):995–7.
23. Cramon-Taubadel N, Frazier BC, Lahr MM. The problem of assessing landmark error in geometric morphometrics: theory, methods, and modifications. Am J Phys Anthropol 2007;134(1):24–35.
24. Ross AH, Williams S. Testing repeatability and error of coordinate landmark data acquired from crania. J Forensic Sci 2008;53(4):782–5.
25. Butler J, Kline M, Coble M. NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): variation observed and lessons learned. Forensic Sci Int Genet 2018;37:81–94.
26. Dror IE, Hampikian G. Subjectivity and bias in forensic DNA mixture interpretation. Sci Justice 2011;51(4):204–8.
27. Barrio PA, Crespillo M, Luque JA, Aler M, Baeza-Richer C, Baldassarri L, et al. GHEP-ISFG collaborative exercise on mixture profiles (GHEP-MIX06). Reporting conclusions: results and evaluation. Forensic Sci Int Genet 2018;35:156–63.
28. Bradley MJ, Keagy RL, Lowe PC, Rickenbach MP, Wright DM, LeBeau MA. A validation study for duct tape end matches. J Forensic Sci 2006;51(3):504–8.
29. Tarone AM, Foran DR. Generalized additive models and Lucilia sericata growth: assessing confidence intervals and error rates in forensic entomology. J Forensic Sci 2008;53(4):942–8.
30. Yuen SKY, Taylor MC, Owens G, Elliot DA. The reliability of swipe/wipe classification and directionality determination methods in bloodstain pattern analysis. J Forensic Sci 2017;62(4):1037–42.
31. Page M, Taylor Blenkin M. Expert interpretation of bitemark injuries – a contemporary qualitative study. J Forensic Sci 2013;58(3):664–72.
32. Dama N, Forgie A, Manica S, Revie G. Exploring the degrees of distortion in simulated human bite marks. Int J Legal Med 2019. https://doi.org/10.1007/s00414-019-02163-5
33. Raymond G, Miller P, Bush P, Bush R, Dorion RB, Bush M. Uniqueness of the dentition as impressed in human skin: a cadaver model. J Forensic Sci 2009;54(4):909–14.
34. Alberink I, Ruifrok A. Repeatability and reproducibility of earprint acquisition. J Forensic Sci 2008;53(2):325–30.
35. LaPorte GM, Stephens JC, Beuchel AK. The examination of commercial printing defects to assess common origin, batch variation, and error rate. J Forensic Sci 2010;55(1):136–40.
36. Kelley S, Gardner BO, Murrie DC, Pan K, Kafadar K. How do latent print examiners perceive proficiency testing? An analysis of examiner perceptions, performance, and print quality. Sci Justice 2020;60(2):120–7.
37. Earwaker H, Nakhaeizadeh S, Smith NM, Morgan R. A cultural change to enable improved decision-making in forensic science: a six phased approach. Sci Justice 2020;60(1):9–19.
38. Dror IE, Langenburg G. "Cannot decide": the fine line between appropriate inconclusive determinations vs. unjustifiably deciding not to decide. J Forensic Sci 2019;64(1):10–5.
39. Cole SA. More than zero: accounting for error in latent fingerprint identification. J Crim Law Criminol 2005;95(3):985–1078.
40. Thomas EJ. The harms of promoting 'Zero Harm'. BMJ Qual Saf 2020;29(1):4–6.
41. Biedermann A, Bozza S, Taroni F, Garbolino P. A formal approach to qualifying and quantifying the 'goodness' of forensic identification decisions. Law Probab Risk 2018;17(4):295–310.
42. Pierce ML, Cook LJ. Development and implementation of an effective blind proficiency testing program. J Forensic Sci 2020;https://doi.org/10.1111/1556-4029.14269
43. Kukucka J. People who live in ivory towers shouldn't throw stones: a refutation of Curley, et al. Forensic Sci Int 2020;2:110–3.
44. Orne MT. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. Am Psychol 1962;17(11):776–83..
45. Paulhus DL. Measurement and control of response biases. In: Robinson JP, Shaver PR, Wrightsman LS, editors. Measure of personality and social psychological attitudes. San Diego, CA: Academic Press, 1991; 17–51.
46. Dror IE. Biases in forensic experts. Science 2018;360(6386):243.
47. Jeanguenat AM, Dror IE. Human factors effecting forensic decision making: workplace stress and wellbeing. J Forensic Sci 2018;63(1): 258–61.
48. Dror IE. A hierarchy of expert performance (HEP). J Appl Res Mem Cogn 2016;5(2):121–7.
49. Daubert v. Merrill Dow Pharmaceuticals, Inc. 509 U.S. 579 (1993).
50. National Commission on Forensic Science. Ensuring that forensic analysis is based upon task-relevant information. Washington, DC: National Commission on Forensic Science, 2016.