Forensic Science International: Synergy 2 (2020) 703–704

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: https://www.journals.elsevier.com/forensic-science-international-synergy/

Continued confusion about inconclusives and error rates: Reply to Weller and Morris

We are happy to see a Commentary [1] about our article "(Mis) use of Scientific Measurements in Forensic Science" [2]. However, the Commentary suffers from three major errors that undermine the criticisms it levels.

First, the Commentary criticizes our "apparent insistence that every call be regarded as either "correct" or an "error"." This reflects a confusion about the *status* of the decision (correct or incorrect; true or false) with the *actual* decision itself. The *actual* decision -whether it is identification, exclusion, or inconclusive– can have the *status* of being either correct or incorrect. For example, the status of an identification decision is correct when called in reference to a same-source comparison that has sufficient detail to justify an identification, otherwise the identification decision is incorrect. The same logic applies to exclusion decisions, and inconclusive decisions when an appropriate experimental design is utilized (see [2]).

Second, the Commentary rejects our proposal to include a critical category of inconclusive *evidence* in error rate studies, and states that "We must deal with the fact that ground truth always has two categories." From a metaphysical perspective, the authors are correct: two items either do –or do not– originate from the same source. But the relevant question is whether sufficient quality or quantity of information exists in the evidence to make a source determination. For example, consider a cartridge casing that was discharged on a freeway, run over by dozens of vehicles, and then swept into a sewer for months to further degrade; that casing may not contain any information pertinent to making a source determination, therefore it is inconclusive evidence. In such cases, an inconclusive decision would be the only correct decision, despite the fact that the casing in question either did –or did not– originate from the same source as the comparison casing. Put differently, the relevant question is not the ontological or metaphysical perspective of *what exists*, but the epistemological perspective of *what we can know*, what we can justifiably conclude given the information available, i.e., whether there is sufficient quality and quantity of information that justifies making a source determination [3]. Otherwise, the *evidence* is inconclusive.

Determining whether evidence does –or does not– have sufficient quality and quantity of information is challenging but not intractable, as we noted in our article [2]. One approach we suggested to overcome the challenge was to have a panel of independent experts predetermine which items lack sufficient quality and quantity of information, and would thus be deemed inconclusive evidence. The Commentary purports to demonstrate how this approach is "flawed" by describing a hypothetical study in which a panel of experts predetermine that a set "*should* be called inconclusive" (emphasis in original). In this hypothetical study the data reveal that the majority of participants nevertheless determined this evidence as an identification, and sure enough, according to the Commentary's hypothetical case, this evidence was indeed from the same source. The Commentary notes that "under the authors' proposed paradigm, these test participants' answers would be erroneous, despite the majority being factually correct when considering the fundamental ground truth: these are same-source comparisons."

The problem with this hypothetical is that the majority of participants contradict the independent experts who predetermined the set "*should* be called inconclusive" due to insufficient quality or quantity of information. Either the evidence is inconclusive or can justifiably be determined as same source, but it cannot simultaneously be both, as the Commentary suggests. Moreover, imagine the participants flipping a coin to determine whether the evidence is an identification or an exclusion; it is irrelevant whether that determination was "factually correct" because the issue is about whether there is a justified basis for drawing the conclusion—and here the independent experts predetermined such as basis does not exist and therefore that the evidence is inconclusive.

Taking the Commentary's own example, let us accept their hypothetical study and data: A piece of evidence is determined by examiners where the "majority being factually correct when considering the fundamental ground truth: these are same-source comparisons." If we accept in this example that the majority are correct and justified in calling it an identification, that *must* entail that all the examiners who concluded that the same evidence is inconclusive are in error and must be included in the error rate calculation as errors. One cannot have it both ways, as the Commentary suggests. If evidence is from the same source, and the majority of examiners in the study are justifiably determining it as an identification, then those examiners who do not call it as an identification and determine it as something else are making an error.

In our article [2] we noted that the use of a panel of independent experts has limitations and we suggested alternative ways to deal with the challenge of determining whether or not evidence is inconclusive. We are open to considering additional possible ways to determining which evidence is inconclusive. However, continuing to deny the existence of the category of inconclusive evidence will stymie productive conversation regarding this important category of evidence, and it leaves the results of error rate studies ambiguous since there is no way to assess whether inconclusive decisions are correct or incorrect.

Third, the Commentary [1] proposes different ways to calculate "meaningful performance metrics" all of which leave inconclusive responses ("I") in the denominator when calculating the error

DOI of original article: https://doi.org/10.1016/j.fsisyn.2020.10.004.

https://doi.org/10.1016/j.fsisyn.2020.10.005

rate. However, as pointed out in our article [2], including inconclusive decisions in the denominator of the error rate calculation has a *de facto* effect of counting inconclusive responses as correct (hence artificially reducing the error rate). The Commentary does not provide any justification for treating all inconclusive responses as correct. Furthermore, as we noted, it cannot be that *all* examiners are correct when they reach *different* conclusions on the same evidence. Worse than that is when the same examiner looking at the same evidence, reaches a different conclusion each time, i.e., contradicting themselves [4]. This faulty approach of counting different/contradicting decisions as all correct has plagued error rate studies [2,4], and is being perpetuated by the proposed "mperformance metrics" in the Commentary.

As we explained in our article [2], the central problem in exisiting error rate studies is that the experimental design has only two categories of *evidence* but three *response* options. This entails three possible ways to count inconclusive *responses* in the studies: 1. Count all inconclusive responses as correct (i.e., include them in the denominator of error rate calculations); 2. Count all inconclusive responses as incorrect (i.e., include them in the numerator of error rate calculations); 3. Drop all inconclusive responses from error rate calculations altogether (i.e., do not include them in either the numerator or denominator of error rate calculations). These approaches produce highly discrepant results. For example, the study by Baldwin, Bajic, Morris and Zamzow [5] (Morris is an author of that study and also an author of the Commentary) used option 1; by counting the inconclusive responses as correct, the study reported a false positive error rate of 1.0%. PCAST [6] rejected their approach and used option 3; by dropping all the inconclusive responses from the calculations altogether, PCAST reported a false positive error rate of 1.5% (50% higher than the error rate reported in the Baldwin et al. study, see PCAST Table 2, page 111). Option 2, counting all inconclusive responses as errors, would give a resulting error rate of 35%. Thus, the error rates range from 1% to 35% depending on how one counts the inconclusive responses within this single study. This is an artifact of not being able to ascertain which inconclusive responses are correct and which are incorrect.

A way out of this quagmire is to use a study design that captures the three categories of *evidence* (which also reflects the evidence in casework). By including inconclusive evidence, one can determine when inconclusive *responses* are correct or incorrect (quite simple: does the decision match the evidence). That is exactly what we proposed in our article [2]. The Commentary fails to rebut this proposal nor does it offer a tenable solution to the current unacceptable and misleading approach to calculating error rates.

Declaration of competing interest

No conflict of interest.

References

[1] T.J. Weller, M.D. Morris, Commentary on: I. Dror, N Scurich "(Mis)use of scientific measurements in forensic science", Forensic Sci. Int.: Synergy (2020).
[2] I.E. Dror, N. Scurich, (Mis)use of scientific measurements in forensic science, Forensic Sci. Int.: Synergy, https://doi.org/10.1016/j.fsisyn.2020.08.006.
[3] I.E. Dror, G. Langenburg, Cannot Decide": the fine line between appropriate inconclusive determinations VS. unjustifiably deciding not to decide, J. Forensic Sci. 64 (1) (2019) 1–15, https://doi.org/10.1111/1556-4029.13854.
[4] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners, PLoS 7 (2012), e32800.
[5] D.P. Baldwin, S.J. Bajic, M. Morris, D. Zamzow, A Study of False-Positive and False-Negative Error Rate in Cartridge Case Comparisons, Ames Laboratory, Defense Forensic Science Center, 2014. Technical Report # IS-5207, available a, https://www.ncjrs.gov/pdffiles1/nij/249874.pdf.
[6] PCAST, President's Council of Advisors on science and Technology report on forensic science, in: Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods, 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

Nicholas Scurich[a,*], Itiel E. Dror[b]
[a] *University of California, 4312 Social and Behavioral Sciences Gateway, Irvine, CA, 92697, USA*
[b] *University College London (UCL), 35 Tavistock Square, London, WC1H 9EZ, United Kingdom*

[*] Corresponding author.
E-mail addresses: nscurich@uci.edu (N. Scurich), i.dror@ucl.ac.uk (I.E. Dror).