Forensic Science International: Synergy 5 (2022) 100273

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy

Inconclusives, errors, and error rates in forensic firearms analysis:Three statistical perspectives

Alan H. Dorfman [a,*], Richard Valliant [b]

[a] National Center for Health Statistics (retired), Bethesda, MD, 20814-1345, USA
[b] Universities of Michigan & Maryland, USA

ARTICLE INFO

Keywords:
Likelihood ratio
Virtual comparison microscopy
Nonresponse
Cognitive bias
Test-blind

ABSTRACT

Error rates that have been published in recent open black box studies of forensic firearms examiner performance have been very low, typically below one percent. These low error rates have been challenged, however, as not properly taking into account one of the categories, “Inconclusive”, that examiners can reach in comparing a pair of bullets or cartridges. These challenges have themselves been challenged; how to consider the inconclusives and their effect on error rates is currently a matter of sharp debate.

We review several viewpoints that have been put forth, and then examine the impact of inconclusives on error rates from three fresh statistical perspectives: (a) an ideal perspective using objective measurements combined with statistical algorithms, (b) basic sampling theory and practice, and (c) standards of experimental design in human studies. Our conclusions vary with the perspective:

(a) inconclusives can be simple errors (or, on the other hand, simply correct or at least well justified);

(b) inconclusives need not be counted as errors to bring into doubt assessments of error rates;

(c) inconclusives are potential errors, more explicitly, inconclusives in studies are not necessarily the equivalent of inconclusives in casework and can mask potential errors in casework.

From all these perspectives, it is impossible to simply read out trustworthy estimates of error rates from those studies which have been carried out to date. At most, one can put reasonable bounds on the potential error rates. These are much larger than the nominal rates reported in the studies.

To get straightforward, sound estimates of error rates requires a challenging but critical improvement to the design of firearms studies. A proper study—one in which inconclusives are not potential errors, and which yields direct, sound estimates of error rates—will require new objective measures or blind proficiency testing embedded in ordinary casework.

1. Background

1.1. Inconclusives in casework

A forensic firearms examiner, when comparing two bullets or two cartridge cases, for example one an “unknown” or “questioned” (from crime scene) and the other a “known” (from a suspect weapon), will in the U.S. ordinarily follow The Association of Firearms and Toolmarks Examiners (AFTE) [1] or the Uniform Language for Testimony and Reports (ULTR) guidelines issued by the Department of Justice [2], and arrive at one of three major conclusions: “identification”—the bullets came from the same gun, “elimination”—the bullets came from different guns, or “inconclusive”—the bullets might have come from the same gun, might have come from different guns. (To keep expressions simple, we will frequently speak only of bullets, implying cartridge cases as well; context should make this clear.) The AFTE criteria [1] further divide the Inconclusives into three sub-categories labeled A, B and C, where A leans to without reaching an Identification, C leans to without reaching an Elimination, and B is neutral. It is not clear what roles these play in casework. Different examiners appear to invoke the inconclusive category in different ways ([3], p. 7). There is also another category, “Unsuitable”, said of a bullet when its markings are so obscure as to forestall any comparison.

It may be noted that the three major categories are not on a par. While identification and elimination refer to a gun or guns which shot the bullets, “inconclusive” refers to the status of the examiner’s

[*] Corresponding author.
E-mail addresses: dorfmans@erols.com (A.H. Dorfman), valliant@umich.edu (R. Valliant).

https://doi.org/10.1016/j.fsisyn.2022.100273
Received 24 April 2022; Received in revised form 22 May 2022; Accepted 26 May 2022
Available online 8 June 2022

impressions: he/she cannot tell whether or not the bullets had a common source. (The corresponding category of the bullets might be termed the “ambiguous” or “insufficiently informative”.) We can put this in different words: an identification or elimination speak to a *past* event concerning the bullets (*did* or *did not* come from the same gun) while inconclusive speaks to the *present* state of the examiner’s mind and the bullets before him or her.

Now the bullets either did or did not come from the same gun. To conclude they did when they didn’t is an error—a false positive. To conclude they did not when they did is an error—a false negative.

Declaring “inconclusive”, the examiner fails to catch the truth. Is it an error?

It is clearly some sort of *failure*, the inability to land on the historical truth. And, in casework, it has consequences: an “inconclusive” when the bullets came from the same gun will fail to contribute to the prosecution of a suspect, when from different guns, to the defense of a suspect. It may be the best the examiner can do in the face of the evidence, and then seems simply unfortunate rather than in any sense erroneous. But there can be exceptions; an examiner might on a bad day miss clear signs of identification or elimination. Some examiners, guided by local laboratory policy, will rely on “inconclusive” when differences in markings call for elimination; this does seem to be a questionable practice, with the non-trivial consequence that evidence possibly useful to the defense is denied. Dror and Langenburg [4] suggest that forensic analysts sometimes lean on “inconclusive” simply to avoid a firm decision. However, in this paper, we premise that by and large inconclusives in firearms casework reflect the limits of the methods and material, and should not be regarded as errors, that inconclusives in casework are analogous to a medical diagnostic that comes up borderline and fails to indicate whether or not the patient has a suspected condition.

1.2. Review of inconclusives in firearms identification studies

Early studies aiming to assess the accuracy of forensic firearms examination focused on seeing whether identifications were done correctly (e.g., Hamby et al. [5], and references therein). These “closed” studies presented an examiner with two sets of bullets, and asked the examiner to match up each bullet in the one set with the bullets in the other set that came from the same gun—“set-to-set” studies. For each bullet/cartridge in the “unknown” set, there was always at least one match in the “known” set. Inconclusives were not an issue. For example, in the Hamby et al. study, only 8 of 7605 “unknown” bullets were judged unsuitable or unmatchable to one of a set of “known” bullets ([5], p. 107). Error rates hover near zero in these studies.

Prompted by a 2009 National Academy of Science report [6] that called for further studies of firearms identification, implying that what had gone before was not satisfactory, Knapp and Garvin [7] carried out a small “open” study in which not all bullets presented could be paired with another bullet from the same gun, probably the first such investigation. Unpublished, the study has gone largely unnoticed, except for a summary in a book on firearms identification by Ron Nichols ([8], p. 129). Nichols notes an overall error rate of 5.1%, a good deal larger than what had been previously seen. He puts this down to “test-taking bias”—many of the examiners, including those with years of experience, were assuming the study was closed, *per* earlier studies, and this preconception biased them in the direction of Identification—evidence that even experienced examiners can be influenced by what they expect to see.

Table 1
Knapp-Garvin (2012) study results.

	Identification	Inconclusive	Exclusion	Source Total
Same Source	316	?	1	?
Different Source	21	?	32	?
Conclusion Total	337	60	33	430

? Indicate where a cell count was not provided in presentation.

The 5.1%, however, does not tell the whole story. Table 1 collects the results reported in the 2012 AFTE presentation by Knapp and Garvin. The 5.1% is the number of clear errors over the total number of comparisons: (21 + 1)/430 = 5.1%. But, unlike in the previous closed studies, there are now a large percentage of inconclusives: 60/430 = 14%. If we remove them from the calculations, the overall error rate would be 22/370 = 5.95%. Mostly this comprises false positives—identifications when the bullets came from different guns—and the false positive rate ignoring the inconclusives is 21/(21 + 32) = 39.6%. So, there is a striking difference in results between this (unexpectedly) open study and previous closed studies.

The NAS report was reinforced by a 2016 Report to the President—the PCAST Report [9]—which included a review of almost all forensic firearms studies to that date (the Presidential Commission seems not to have been aware of the Knapp-Garvin study) and found all studies wanting with one exception—the Ames Study [3]. This large study was not only open but pairwise, making it harder for one comparison to shed light on another, and easier to track exactly how many comparisons were being made. Instead of comparing one large set of bullets overall to another set, it required examiners to perform a succession of comparisons of “questioned” to “known“ cartridges. In these conditions a large percentage of comparisons were inconclusives, comprising 23% of all comparisons, and 34% of different source comparisons ([3], pp 15–16).

A successor study initiated by the FBI and carried out by the Ames Laboratory [10] [1] (“Ames-FBI”) likewise exemplifies a large open, pairwise study, and included both bullets and cartridges. Table 2 ([10], p. 34; [12], p. 10) gives the results for “Round 1” of the study, aimed at assessing accuracy. For bullets, there were inconclusive decisions in 51% of all the comparisons, 65% among the different sourced; for cartridges, 42% and 51%.

This amount of inconclusives can lead to no small difference in assessments of error rates. If, like the authors of the study, we include the inconclusives among the number of comparisons and regard them as neutral non-errors, then the error rates are very small. For example, the authors calculate the false positive rate for bullets, using the second row of the Bullets section of Table 2, as F-Pos = 100% x Identification / (Identification + Inconclusive-A + Inconclusive-B + Inconclusive-C + Elimination) = 100% x 20 / 2842 = **0.704%** ([10], p. 34; [12], p. 10)

By contrast, if the inconclusives are regarded as potential errors, as argued by Dror and Scurich [14], discussed below, then the potential error rate for different source bullets is (20 + 268 + 848 + 743)/2842 = **66.1%**. For forensic purposes, *potential* error rates can carry the same weight as error rates *per se* (Daubert [15], pp. 580, 594).

It is clear that a well-founded characterization of inconclusives is critical for assessing the size of error rates estimated from the forensic firearms studies.

1.3. Review of characterizations of inconclusives

Hofmann et al. [16] offer four distinct Options for looking at Inconclusives, give a rationale for each, and, for three of the options, calculate the corresponding error rates in several recent studies,

[1] Reference [10] is the full October 2020 127-page Report of the Ames laboratory to the Federal Bureau of Investigation, comprehensively detailing data and analysis estimating accuracy, repeatability, and reproducibility *inter alia* of forensic firearms examinations. We will reference it despite its not being available in the public domain as of this writing. It was released to the public in early 2021 and then withdrawn. Before being withdrawn, it circulated widely enough to have been put into evidence in several court cases. We include it therefore as being important and available to some readers but limit ourselves to material also cited in publicly available (derived) sources, one or other of references [11–13].

Table 2
Conclusions in Round 1 of Ames-FBI Study estimating Accuracy.

Bullets						
	Identification	InconclusiveA	InconclusiveB	InconclusiveC	Elimination	Source Total
Same Source	1076	127	125	36	41	1405
Different Source	20	268	848	745	961	2842
Conclusion Total	1096	395	973	781	1002	4247
Cartridge Cases						
	Identification	InconclusiveA	InconclusiveB	InconclusiveC	Elimination	Source Total
Same Source	1056	177	140	22	25	1420
Different Source	26	177	637	620	1375	2835
Conclusion Total	1082	354	777	642	1400	4255

Based on ([10], Table V, p. 34; [12], Table 2, p. 10).

including the original Ames study.

Option 1. Inconclusives as *ignorable*. Thus, setting them aside in any calculation of error rates. Hofmann et al. suggest, however, that inconclusives have real consequences in practice and cannot be ignored. It is the one option which they bypass in their tabulations.

Option 2. Inconclusives as *correct*, whether in context of same source or different source bullets. They describe this as a reasonable position under prevailing AFTE guidelines and refer to the corresponding error rate as *examiner error*.

Option 3. Inconclusives as *errors*. Either the bullets came from the same source or a different source; to fail to reach the correct conclusion is an error. The corresponding error rate they refer to as *process error*.

Option 4. Inconclusives as *eliminations.* In this interpretation they are regarded as failures to identify, and so an error only under same source conditions.

The error rates as calculated in Ames-FBI [10,12], and virtually all firearms studies, are mathematically equivalent to adopting Option 2 ("correct"). The reader is referred to Hofmann et al. [16] for the error rates under the different options for a multitude of recent studies.

The error rates under the different options for the data in Table 2 above are summarized in Table 3.

Dror and Scurich [14] argue that in a well-designed forensic study, *some* of the inconclusives will be errors and *some* will be correct. Which are which will have an objective foundation, based on the condition of the bullets or cartridges presented to the study's examiners. Some pairs of bullets are sufficiently unambiguous in their markings that "identification" or "elimination" is well justified and declaring anything else but the appropriate choice, including "inconclusive", is an error. Some are sufficiently vague or ambiguous in their markings that declaring anything but "inconclusive" is wrong. A proper study will assess examiners' decisions on this objective basis. Dror and Scurich grant that getting this objective basis is difficult, but as an approximation to it they suggest workarounds of either pre-study expert assessment of the bullets or using a "majority rules" of examiners in the study looking at the same items.

Where the designers of the study have made no attempt to establish this basis, that is, establish a class of objective inconclusives, all one can say is that in all likelihood some of the inconclusive decisions are correct, others are not. If all of the inconclusives were decided correctly, then the error rates are small; if all were decided incorrectly, the error rates are huge. All we can then properly speak of is the *potential* error rates, which can be assumed to lie somewhere between the minimum and the maximum, i.e. somewhere between Option 2 and Option 3. Thus, for example, Ames-FBI different source potential error rate for bullets lies between 0.70% and 66.2%. Since, for judicial purposes, the scientific burden is to prove small rates, one must be concerned about the potentially large rates.

Table 3
Error Rates for Ames-FBI Study using Hofmann et al. Classification.

Bullets				
Inconclusive as …	*Ignorable*	*Correct*	*Error*	*Elimination*
same source	3.67%	2.92%	23.42%	23.42%
different source	2.04%	0.70%	66.19%	0.70%
Cartridge Cases				
Inconclusive as …	*Ignorable*	*Correct*	*Error*	*Elimination*
same source	2.31%	1.76%	25.63%	25.63%
different source	1.86%	0.92%	51.50%	0.92%

Dror and Scurich's perspective has been challenged both at its root—whether any ground truth besides the *de facto* source of the bullets can be established—and because of the paradoxical results that can arise from the use of pre-assessment or majority rule [17,18]. "While it is difficult to disagree with Dror and Scurich …. that current practices for processing 'inconclusives' are unsatisfactory, and prone to adversely affect standard procedures for computing error rates, the proposed remedies only compound and shift the problem."—Biedermann and Kotsoglou ([18], p. 2).

We here take a fresh look at the inconclusives from three statistical perspectives: (a) an ideal perspective using objective measurements, statistical algorithms and likelihood theory, (b) fundamentals of survey sampling, and (c) fundamentals of experimental design in human studies. The conclusions we arrive at complement each other: (a) with sufficient objectively established background, inconclusive decisions can be divided into those that are correct and those that are errors, (b) the presence of inconclusives in standard firearms forensic studies brings into question the accuracy of estimates of error rates, and (c) inconclusives are potential errors and open the door to high potential error rates in casework.

2. Three statistical perspectives

2.1. An ideal: mechanical firearms examination

Consider some class of guns, delimited for example by brand, model, period of manufacture. Suppose, having examined many bullets shot from many guns in the class, having taken a variety of measurements on each bullet, we have been able by statistical means to get the frequency of occurrence of an overall measure of closeness "c" of markings between two bullets when (a) the two bullets came from the same gun and (b) from different guns. An idealized version of such frequencies is given in Fig. 1, where the bold curve refers to the frequency of c when the bullets came from the same gun and the dashed curve represents the frequency, when from different guns.

Now suppose we are presented with two bullets from a gun or guns belonging to the above class, a "questioned" (from crime scene) and a "known" (from suspect's weapon) and we find their measure of closeness is c_{kq}. What we can or should conclude will depend on the relative sizes of the frequencies at that particular value of c. The more c_{kq} is to the right, the greater the relative probability of the bullets having arisen from the same source, and the more c_{kq} to the left, the greater the relative probability of different source.

In the figure it happens that the two curves intersect at $C = 0.535$ so if $c_{kq} > C$, that is evidence suggesting the bullets came from the same (suspect's) gun, and the farther to the right is c_{kq}, the stronger that evidence. Likewise, the farther below C, the stronger the evidence of an

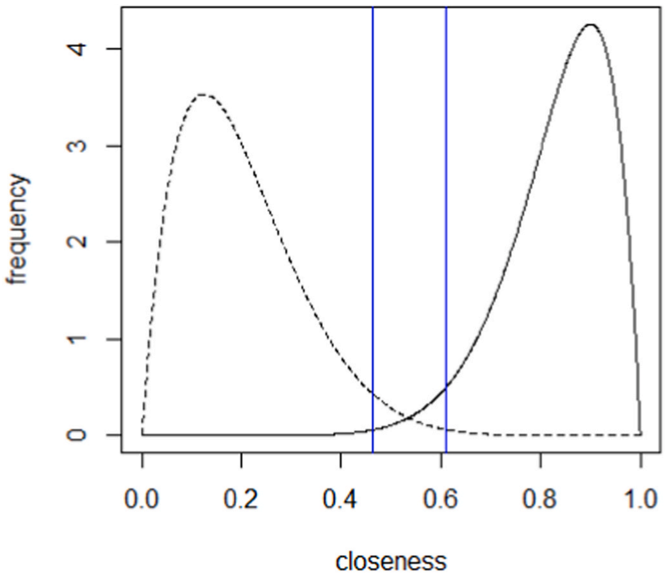

Fig. 1. *Hypothetical frequency of measure of closeness of bullet pairs from same and different guns.*

"elimination" (the unknown did not come from suspect's gun).

What we are being guided by is the relative heights of the two curves. The ratio of these heights at any c is called the *likelihood ratio* at c. For example, it turns out the likelihood ratio at $c = 0.8$ is about 4000 in favor of "same gun" at $c = 0.275$, about 3000 in favor of "guns are different". In both cases, provided our measurements have all been sound, the evidence is clear-cut.

If it happened that comparison of our known and questioned bullets yielded $c_{kq} = C = 0.535$, then neither option is favored, and this case justifies the conclusion that the results are neutral, leaning neither way, ambiguous—"inconclusive".

The chances of getting this exact value are small. It seems reasonable to extend "inconclusive" to values *near* C. Even if c were a bit to the right of C, the chances of such a result, even if the bullets did arise from different guns, would not be remote.

One way to delimit the inconclusive region is to take the c's where the likelihood ratio is *low* in either direction. How low, in the context of firearms, is a matter for further discussion, but a general rule suggested in a well-known text on the use of likelihood ratios as evidence is to use a likelihood ratio of 8 as cutoff ([19], sections 1.6.1, 4.4); ratios less than that generally provide weak evidence. In our artificial example depicted in the figure, the bounds of the "inconclusive" region, marked by the vertical lines, are from $c = 0.46$ to $c = 0.61$. By this criterion, it would be fitting to declare "inconclusive" if c fell in that range.

If we grant the above—all the background information that went into getting the measure of closeness c and its distributions under the two possibilities of same or different source, and agreement on the "neutral region"—then it would be procedurally erroneous to declare an identification when $c = 0.55$. Likewise it would be procedurally erroneous to declare an "inconclusive" rather than an elimination when $c = 0.4$. In the forensic firearms context, Riva and Champod [20] suggest basing the inconclusive boundaries on the probabilities of false classification they entail. We pass over details here.

2.1.1. *Relevance to Dror and Scurich experimental design*

Under the above ideal, mechanical scheme, there are true and false inconclusives. Someone reporting an inconclusive when the mechanical system shows $c = 0.30$ is committing an error.

As noted above, Dror and Scurich posit that a proper study of firearms examiners' skills requires a ground truth that includes not only same gun, different gun, but also true inconclusives ([14] Fig. 1, p. 3). The mechanical scheme we have described, were it available for the guns included in the study, would provide an objective basis for a ground truth of inconclusive, and provide a less personal alternative to the approaches requiring pre-assessment by other examiners, or relying on majority rule. It is to be noted that this scheme could be applied retroactively to studies already completed, once there was sufficient technical background and statistical calculation to get the likelihood curves for the particular guns and ammunition in the study, leaving open till then what the objective error rates are for Ames, etc.

2.1.2. *Feasibility*

Although we have spoken of the above mechanical scheme as an "ideal", it should be recognized that there has been a good deal of research into developing such schemes, using 3D-virtual computer microscopy (VCM) and "machine learning" (sophisticated statistical approaches to classifying data), for example, Carriquiry et al. (2019) [21], Riva et al. (2020) [22] and other works they reference. These approaches show great promise, and, while not yet ready for full use in forensic laboratories, may well be adopted in the years ahead. One barrier to overcome is the sheer task of acquiring the massive background data required to construct the frequencies of measures of closeness for the many classes of firearms that exist (and continue to evolve), with possible further breakdowns by ammunition used.

It is an open question whether algorithms applied to data from virtual microscopy could eventually altogether replace traditional "subjective" judgments of firearms examiners. Mattijssien et al. [23] give results of standard firearms examinations against a background of results of virtual microscopy-cum-algorithms. The study was narrowly focused on firing pin aperture shear marks generated by a single brand/model. The algorithm they used got near perfect results: there were *zero* inconclusives (in effect, the two frequency curves had zero overlap) and an overall error rate of 1.25%. Firearms examiners looking at a subset of the totality of comparisons found 28.2% of them to be inconclusives. By the Dror-Scurich standard, *all* the examiner inconclusives would be errors in this case, yielding a 31.8% overall error rate, higher for different source. Because of the specialized focus of the study (ignoring all but one of breech face characteristics), we hesitate to draw any broad implications.

2.1.3. *Use of VCM for imaging*

The 3D imaging technology applied to bullets and cartridges produces sophisticated computer data capturing their topography, which can actually be used in *two* ways. We have been emphasizing their digitized use, the combining of the data to form measures of closeness that, once satisfactory algorithms are in place, can be used to yield objective evidence about pairs of bullets of interest.

The data can also be converted into computer images much like the standard 2D images examiners observe through comparison microscopes. The 3D computer images have several advantages over the 2D; for example, better control of lighting and increased detail. Especially noteworthy is they readily allow for examiners to examine pairs of bullets/cartridges without requiring their physical presence. A dozen examiners widely spaced out geographically can examine the same cartridge pair contemporaneously. This "analogue" use of the 3D data is already being made use of by some forensic laboratories, at least for cartridge cases. Lilien et al. [24] list several applications that are already feasible, including validation studies, proficiency testing, and blind verification. If adopted by a large network of forensic laboratories, one can envisage use of this technology in large scale blind studies (see end of section 2.3 below).

2.2. *Sampling theory*

The goal of forensic firearms studies is to estimate the error rates in actual casework. Casework consists of the large number of comparisons made by expert examiners on a daily basis, as they examine bullets and cartridges expelled from guns of interest or seek to determine the number of guns used at a crime scene or whether the same gun was used at two different crime scenes. The error rates in forensic firearms studies are thus always *estimates* of the actual error rates and the studies that support these estimates can be thought of as *samples* from the vast population of comparisons in fieldwork.

The studies are samples in the same sense that pre-election polls are samples from the votes cast in the election. Just as the polls do not actually sample from votes which are not yet cast, but instead seek a surrogate ("if the election were held today how would you vote?"), so too the examinations in the studies are not drawn from actual casework, where ground truth as to source is not known, but are instead intended to be *like* casework, only with ground truth known.

For these estimates to be trustworthy, an experimenter carrying out a study estimating firearms comparisons accuracy has at a minimum to keep in mind the basics of sampling theory. There are two bedrock concerns in any sample survey: *how* the sample is selected, and minimizing or accounting for *nonresponse*, the sometimes failure to acquire information from those sampled.

With regard to the sampling itself, what the experimenter-sampler wants to do is to control the method of selection so that the sample will be *representative*—a sort of miniature of the population, having a similar spread of relevant characteristics. There are standard ways of selecting a good sample, typically involving some species of probability (random) sampling [25]. Random sampling is not necessarily a condition for getting a representative sample, but alternatives need to be fairly sophisticated [26]. In particular, allowing the examiners in the sample to be *self-selected*—a species of *convenience* sample—brings on a suspicion of biased estimates, unless special measures can be and are taken at the estimation stage [27]. There is also the enormous difficulty of getting a representative sample of firearms and ammunition [28,29]; see sub-section 2.2.2 below.

However well done the sampling, there then remains the task of taking into account the almost inevitable presence of nonresponse, the fact that not all the information sought is made available by those sampled. There are *two* kinds of nonresponse: *unit* nonresponse, and *item* nonresponse ([30] p. 559; [31]). In the case of firearms studies, these would correspond respectively to whether an examiner, once enrolled, stays in the study, and then, among those that do, to whether he or she reports definitive results on a particular comparison.

All three—sampling method, unit nonresponse, item nonresponse—can raise concerns about the reliability of estimates in a forensic firearms study. For example, the Ames-FBI study, like most, possibly all, forensic firearms studies, selected its participants on a volunteer basis ([10] p. 2 [11], p. 11), leading to doubts about whether it is a representative sample. In addition, there was an enormous amount of dropout from the study, leading to a concern about unit nonresponse bias—32% of the 256 examiners receiving their first packets failed to report any results, and another 32% of the 256 dropped out before completing all six mailings: ([10] pp. 15, 17, 31; [11], p. 11, Fig. 5).

In the field of survey sampling, there is not an accepted standard for when a nonresponse rate becomes so high that data are unusable. The key consideration is whether the act of responding is related to what is being measured [45]. Samplers can sometimes make use of sophisticated methods of adjusting for nonresponse ([30], Chapter 15, esp. Section 15.4; [46], Section 13.5). However, the need for such measures is rarely, if ever, recognized in the forensic firearms studies. For example, the Ames-FBI report deprecates the importance of the large dropout rate in the study by suggesting it was in large measure due to examiners' busy schedules ([10] p. 15; [11], p.11 [12]; p. 4). Thus, they tacitly assume that examiners with greater burden make mistakes at no different rate than those in more relaxed environments. There is no serious entertainment of other possible factors that could lead to a difference in proficiency between those who stayed in and those who dropped out.

Inconclusives play a role as suspect item nonresponse. Dror and Langenburg [4] point out the ambiguity of inconclusive decisions: they *can* be a straightforward assessment of the (inadequate) information in the items compared, but they can also be a means of avoiding expressing an opinion as to source. The latter is a distinct possibility in forensic studies, where examiners in difficult comparisons risk pronouncing a conclusion contrary to known ground truth, possibly contributing to doubts "… about processes that they have been doing for a long time and in which they are heavily invested." [28] From the point of view of calculating error rates, one is left in the dark as to whether the comparisons termed inconclusive are equivalent to what would be judged inconclusives in field work or are a means for sidestepping a statement about source (see section 2.2.3 below). Considering that the hypothesis being tested is one of low error rates, it would be improper to ignore the possibility that in at least some cases, inconclusives are instances of unit nonresponse, biasing results. We conclude that, from a sampling perspective, the inconclusives need not be regarded as errors *per se*, in order for their presence to cast doubt on the accuracy of nominal error rates.

2.2.1. *Confidence intervals*

Besides leading to possibly biased estimators, poor sample design and the presence of nonresponse undermine confidence interval construction. This is important to note because taken at face value confidence intervals for error rates might mistakenly be thought to bound the potential error rates. For example, Ames-FBI calculates the upper bound of a 95% confidence interval of the false positive error rate for bullets to be 1.42% ([10], Table VIII, p. 36; [12], Table 5) and does so using a relatively sophisticated model to take into account differences in examiner performance.

Unfortunately, given the sampling weaknesses noted above, the Ames-FBI confidence intervals, despite their sophistication, are not well-grounded. What a confidence interval is intended to convey is the range within which we might expect to find the average of error rates in the whole population of firearms examiners. There are many ways besides simple random sampling to select samples that can yield good population estimates. The opportunistic sampling of Ames-FBI, which we pointed to, is not one of them. If the sample is opportunistic and beset by nonresponse at both the unit level and item level, then it is risky to assume it represents the full population. Distortions can arise. For example, the small minority of poorly performing examiners in the sample might represent a greater proportion in the population. This implies that the Ames-FBI confidence intervals, however well constructed, cannot be taken at face value. In particular, the bounds of the proffered confidence intervals cannot be taken as setting limits on the potential error rates in casework.

2.2.2. *Selection of weapons for study*

With a few exceptions, each of the forensic firearms studies to date focuses on a single firearm. This gives rise to two concerns. The first is the *difficulty of generalizing* results to the population of firearms examinations in general. One cannot reach a conclusion about error rates in the great variety of firearms comparisons in forensic laboratories by focusing on comparisons of bullets or cartridges fired from say 9 mm Ruger pistol barrels, as in the very large Brundage Hamby study (bullets) [32] or the Ames study (cartridges) [3]. The few studies that have carried out comparisons over a variety of guns have displayed marked differences in the ease of coming to correct conclusions [33,34].

The second concern is the *reduction of independence* of comparisons, even in pairwise studies. If an examiner is over and over comparing bullets or cartridge cases from the same brand and model, then he or she can be expected to be picking up nuances along the way. A later comparison will have an advantage over the first. We can expect this to lead to a reduction in sample error rates. Tellingly, the original Ames Report, which emphasized the independence of its comparisons, at one point acknowledges that this is not quite so: "The study was designed with sample sets for comparison that are as independent *as economically feasible* given the cost of firearms and ammunition" ([3], p. 5, emphasis added).

To some extent, the FBI-Ames study attenuated this problem by using two distinct guns for both bullets and cartridge cases. For bullets, the FBI chose to use Beretta Model 92 and Ruger SR9C. They remark that these were chosen for the difficulty of comparisons they generate, so, presumably, most prone to error ([10], pp. 2, 19; [11], pp. 3–4). Beretta and Ruger were among several brands that Bachrach [33] employed in an early 3-D study, and precisely these two brands yielded error rates that were a good deal *lower* than the other brands considered; see Ref. [33] page 51, Fig. 27, and discussion in 4.6.3.1. This of course does not disprove the FBI claim: the models were not the same as used in Ames-FBI and also Bacharch did describe his methodology as a work in progress. But it would be interesting to know the evidence that in the newer models the situation reverses.

2.2.3. *Ambiguity of "inconclusive": evidence from repeatability*

In the Ames-FBI study, there are additional reasons for thinking a fairly large portion of the inconclusives represent item nonresponse rather than "hard core" inconclusives. There is the rather large amount of inconclusives, roughly 50% of comparisons in the case of bullets, and 40% in the case of cartridges, roughly double the already large percentages in the original Ames Study. Possibly, this is due to the greater difficulty attributed to the ammunition and firearms used in Ames-FBI. However there is also the *low repeatability* the study displayed, with examiners bouncing back and forth between categories when re-examining items. The repeatability for bullets was 79.0% and 64.7% for same and different source respectively; for cartridges, these were 75.6% and 62.2%. ([10], pp. 37–45; [13]). So examiners examining the same material twice, disagree with themselves between 20 and 40% of the time. This suggests a fair degree of instability of decision and a kind of fluidity of boundaries between categories. The Ames-FBI repeatability and reproducibility results are discussed in detail in Ref. [35].

2.3. *Experiments with human subjects*

Forensic firearms studies are *experiments on human subjects*, basically aimed at measuring how well the training and experience of firearms examiners enables them to make correct judgments about the source of bullets or cartridge cases. The studies are like the experimental trials in which new medications are tested on human subjects. In both situations, the reactions of subjects are carefully tracked: in the firearms case to the presentation of bullets or cartridge cases, in the medical case to the receipt of the medication being tested or something else.

It is recognized standard procedure that the medical trials be *double-blind*. This means that (1) the subjects are unable to discern whether they have been given the medication or a placebo, and (2) those medical workers administering the medication and observing the results are likewise kept in the dark whether medication or placebo was administered. This practice arises from the longstanding recognition that exposure to such information can affect what the treated experience, and it can also affect the assessments and even observations of those doing the treating [36]. In other words: there is ample evidence that cognitive biases can play a pervasive role in the exercise of human perception and judgment in scientific investigations involving human subjects, and reducing information is a key way to combat that tendency.

The desirability of blinding is acknowledged within the forensic firearms experimental community. For example, "It [the Ames-FBI study] was designed as a true double-blind "black box" investigation, with contact between the participating examiner subjects and the experimental team restricted at all times to both preserve anonymity of the participants and prevent any interactions between participants and investigators that might result in bias." ([12], p. 3). There was careful segregation between those who produced the specimens, those who communicated with firearms subjects, and those who designed and analyzed the experiment ([10], p. 13; [11], Section 2.7). Monson et al. tabulate 18 previous forensic firearms studies, listing only one as not blinded, 11 as single blind and 6 as double blind ([11], Table 1), this despite citing a recent well thought out paper [37] that decries the ambiguity in "double blind" and calls for the clear delineation of who is blinded and with respect to what.

What *none* of these studies are, including Ames-FBI itself, is "test-blind", a term coined by PCAST ([9], p. 58). There is a well-recognized risk of bias, arising from the mere fact the examiners are aware they are being tested. and that something very important—the *bona fides* of their profession—is on the line. Since every "hard error"—false negative and false positive—undermines the appearance of near perfection, it can be expected that examiners who know they are under study will, consciously or unconsciously, lean away from risking "hard errors"—-they can opt for an Inconclusive when there might be some doubt.

This opens the door to the possibility that different source bullets/cartridges deemed inconclusive in a standard non test-blind study would, in the different context of casework, have been judged identification. The concern is heightened by the accumulating evidence that contextual information in casework biases in the opposite, less cautious, direction from the studies (see Ref. [38] and references therein). We noted in section 2.2.3 the low repeatability, i.e. the non-trivial tendency of examiners to change categories on a second examination of given material. We would not expect the amount of shifting between study and casework to be any less than that found within the study.

Blind testing in the key sense—test-blind, in which the firearms examiners do not know they are being tested—has been proposed for a long time. "How can we provide legal decision makers with an empirically based sense of the frequency with which various types of forensic testimony are wrong or misleading? The answer is to conduct methodologically rigorous, blind, external proficiency tests using realistic samples" [39]. Such studies would include the disguised interweaving of fake cases, where true source of bullets/cartridges are known, into the flow of ordinary casework. The 2016 PCAST Report pushes hard on the idea, and, while acknowledging the cost and logistical barriers, envisioned wide adoption of test-blind testing by 2021 ([9], p. 58–59]. Blind proficiency testing does already exist in embryo in some labs [40–42]. However, adoption of measures to remove task-irrelevant information [43,44] in ordinary casework will almost certainly be necessary to assure valid test-blindness. Data on examiner output in casework, e.g. rate of inconclusive responses, should be captured prior to the implementation of blind testing programs to allow researchers to compare data before and after implementation of the program and evaluate the impact of introducing the blind testing program itself. As virtual 3D microscopy gains traction in forensic firearms laboratories (see 2.1.3 above), the wide scale implementation of blind testing, and with it, the ability to carry out large scale test-blind studies should become much more feasible. The establishment of "an independent federal entity, the National Institute of Forensic Science ...." ([6], p. 81) to oversee the studies, although not likely to happen soon, would provide additional assurance of scientific soundness.

In any case, until test-blind studies are implemented, we must regard the forensics firearms studies as yielding inconclusives that are *potential errors*, in the critical sense of masking the potential to be hard errors were the same material presented in casework. It follows that the *potential error rates* are higher, and likely a good deal higher (witness the Knapp-Garvin study), than the nominal rates coming out of forensic firearms studies so far, even those as sophisticated as Ames-FBI.

3. Conclusion

We have considered the question of the impact of inconclusive decisions in firearms forensic studies on estimates of error rates, from three standpoints: (a) ideal objective measures employing virtual 3-D microscopy and standard statistical algorithms, under development, (b) survey sampling fundamentals concerning method of sampling and the issue of nonresponse, and (c) experimental design in studies involving human subjects, with special attention to cognitive bias. From these three standpoints, we conclude, with regard to current and past forensic firearms studies, that, respectively, (a) there can be an objective basis for regarding inconclusives as actual errors, (b) inconclusives as item nonresponse contribute to bias in estimates of error rates, and (c) a decision of inconclusive in a non-test-blind study can mask what would be a mistaken identification or elimination in casework, which together substantially reduce the credibility and reliability of the error rates reported in the studies and the extent to which the reported error rates can be assumed to generalize to real casework.

There are two paths forward to designing studies that could give accurate estimates of error rates in casework. The first lies through the path of employing statistical algorithms applied to data from virtual 3D microscopy to derive objective measures of closeness and likelihood ratios against which to measure examiners' subjective conclusions. The second is to employ appropriate wide scale test-blind studies resting on well conducted blind proficiency testing. Neither is beyond the feasible, and, until one or the other is implemented, sound estimates of error rates are elusive and, in light of the over-abundance of inconclusives, potential error rates must be considered large.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

The first author is a recently retired Senior Mathematical Statistician in the Office of Survey Methods Research at the U.S. Bureau of Labor Statistics (BLS) and Senior Consultant at the U.S. National Center of Health Statistics. His interest in forensic science occurred as a result of his exposure to a case in which a person who had worked at BLS was convicted of a crime based on firearms comparison evidence. The apparent statistical issues involved spurred his current interest in the subject-matter. The views and conclusions contained herein, however, are those of the authors.

References

[1] AFTE, Range of Conclusions, Association of Firearm and Tool Mark Examiners, 2020. https://afte.org/about-us/what-is-afte/afte-range-of-conclusions. (Accessed 17 March 2022).

[2] Department of Justice. Approved ULTR for the forensic firearms/toolmarks discipline – Pattern Examination (effective 8.15.20). https://www.justice.gov/olp/page/file/1284766/download.

[3] D. Baldwin, S. Bajic, M. Morris, D. Zamzow, A Study of False-Positives and False-Negative Error Rates in Cartridge Case Comparisons, Ames Laboratory, 2016. USDOE, Technical Report #IS-5207 ("Ames "), https://www.ojp.gov/pdffiles1/nij/249874.pdf. (Accessed 17 March 2022).

[4] I.E. Dror, G. Langenburg, "Cannot decide": the fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide, J. Forensic Science 64 (2019) 10–15.

[5] J.E. Hamby, D.J. Brundage, J.W. Thorpe, The identification of bullets fired from 10 consecutively rifled 9mm ruger pistol barrels: a research project involving 507 participants from 20 countries, AFTE Journal 41 (2009). Number 2.

[6] NRC, Strengthening Forensic Science in the United States: A Path Forward, National Research Council, National Academy of Sciences, 2009. https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf. (Accessed 17 March 2022).

[7] J. Knapp, A. Garvin, Consecutively Manufactured 25 AUTO F.I.E. Barrels—a validation study, in: AFTE 43rd Annual Training Seminar, 2012. Buffalo, NY, https://afte.org/store/product/training-seminar-dvd-afte-2012-buffalo-ny. (Accessed 17 March 2022).

[8] R. Nichols, Firearm and Toolmark Identification: the Scientific Reliability of the Forensic Science Discipline, Academic Press, 2018.

[9] PCAST, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, President's Council of Advisors on Science and Technology, 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf. (Accessed 17 March 2022).

[10] S. Bajic, L.S. Chumbly, M. Morris, D. Zamzow, Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons, Ames Laboratory-USDOE Technical Report # ISTR-5220 Prepared for the Federal Bureau of Investigation ("Ames II" or "Ames-FBI"), 2020.

[11] K.L. Monson, E.D. Smith, S.J. Bajic, Planning, design and logistics of a decision analysis study: the FBI/Ames study involving forensic firearms examiners, Forensic Science International: Synergy 4 (2022), https://doi.org/10.1016/j.fsisyn.2022.100221.

[12] L. S. Chumbley, M. D. Morris, S. J. Bajic, D. Zamzow, E. Smith, K. Monson, G. Peter. Accuracy, repeatability, and reproducibility of firearm comparisons, Part 1: accuracy (work in progress) https://arxiv.org/ftp/arxiv/papers/2108/2108.04030.pdf ["part 1"] (accessed 17 March 2022).

[13] E. Smith, K.L. Monson, J.L. Stephenson, L.S. Chumbley, S.J. Bajic, M.D. Morris, D. S. Zamzow, The Accuracy, Repeatability, and Reproducibility of Firearms/Toolmarks Comparisons 2022 Firearm and Toolmarks Policy and Practice Forum (Day 4), 17 March 2022. https://forensiccoe.org/2022-firearm-toolmarks-forum.

[14] I. Dror, N. Scurich, (Mis)use of scientific measurements in forensic science, Forensic Science International: Synergy (2020), https://doi.org/10.1016/j.fsisyn.2020.08.006.

[15] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, Me, 1993.

[16] H. Hofmann, A. Carriquiry, S. Vanderplus, Treatment of inconclusives in the AFTE range of conclusions, Law, Probability and Risk 19 (3–4) (2020) 317–364. https://doi.org/10.1093/lpr/mgab002.

[17] T.J. Weller, M.M. Morris, Commentary on: I. Dror, N Scurich "(Mis)use of Scientific Measurements in Forensic Science" Forensic Science International: Synergy, 2020, https://doi.org/10.1016/j.fsisyn.2020.08.006.

[18] A. Biedermann, K.N. Kotsoglou, Forensic science and the principle of excluded middle: 'inconclusive' decisions and the structure of error rate studies, Forensic Science International: Synergy 3 (2021), 100147, https://doi.org/10.1016/j.fsisyn.2021.100147.

[19] R.M. Royall, Statistical Evidence: A Likelihood Paradigm, Chapman and Hall CRC, 1997.

[20] F. Riva, C. Champod, Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases, J Forensic Sci (2014), https://doi.org/10.1111/1556-4029.12382.

[21] A. Carriquiry, H. Hofmann, X.H. Tai, S. Vanderplas, Machine Learning in Forensic Applications, Significance, 2019. https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2019.01252.x.

[22] F. Riva, E.J.A.T. Mattijssen, R. Hermson, P. Pieper, W. Kerkhoff, C. Champod, Comparison and interpretation of impressed marks left by a firearm on cartridge cases – towards an operational implementation of a likelihood ratio based technique, Forensics Science International (2020). https://www.researchgate.net/publication/342074920_Comparison_and_interpretation_of_impressed_marks_left_by_a_firearm_on_cartridge_cases_Towards_an_operational_implementation_of_a_likelihood_ratio_based_technique. (Accessed 17 March 2022).

[23] E.J.A.T. Mattijssen, C.l.M. Witteman, C.E.H. Berger, N.W. Brand, R.D. Stoel, Validity and reliability of forensic firearm examiners, Forensic Science International 307 (February 2020), https://doi.org/10.1016/j.forsciint.2019.110112.

[24] R. Lilien, C. Chapnick, E. Meschke, P. Duez, T. Weller, Expanding the scope and efficiency of 3D surface topography analysis in firearms forensics, in: NIJ Forensic Science Research and Development (R&D) Symposium and Poster Session, March 1-2 2022, 2022. https://forensiccoe.org/2022-nij-forensic-science-rd-symposium/.

[25] S. L Lohr, Sampling Design and Analysis, third ed., Chapman and Hall/CRC, 2021.

[26] R. Valliant, A.H. Dorfman, R.M. Royall, Finite Population Sampling and Inference: A Prediction Approach, Wiley, 2000.

[27] R. Baker, J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, R. Tourangeau, Report of the AAPOR Task Force on Non-probability Sampling, 2013. https://www.aapor.org/aapor_main/media/mainsitefiles/nps_tf_report_final_7_revised_fnl_6_22_13.pdf. (Accessed 17 March 2022).

[28] I.E. Dror, The error in "error rate": why error rates are so needed yet so elusive, Journal of Forensic Sciences 65 (4) (2020) 1034–1039, https://doi.org/10.1111/1556-4029.14435.

[29] C. Spiegelman, W.A. Tobin, Analysis of experiments in forensic firearms/toolmarks practice offered as support for low rates of practice error and claims of inferential certainty, Law, Probability and Risk 12 (2013) 115–133, https://doi.org/10.1093/lpr/mgs028.

[30] C.E. Sarndal, B. Swensson, J. Wretman, Model Assisted Survey Sampling, Springer, 1992.

[31] T. Yan, R. Curtin, The relation between unit nonresponse and item nonresponse: a response continuum perspective, International Journal of Public Opinion Research 22 (Issue 4) (2010) 535–551, https://doi.org/10.1093/ijpor/edq037.

[32] J. Hamby, D. Brundage, N.D.K. Petraco, J. Thorpe, Worldwide study of bullets fired from 10 consecutively rifled 9MM RUGER pistol barrels – analysis of examiner error rates, Journal of Forensic Sciences 64 (2) (2019) 551–557, https://doi.org/10.1111/1556-4029.13916.

[33] B. Bachrach, A Statistical Validation of the Individuality of Guns Using 3D Images of Bullets, Document No.: 213674 Date Received, March 2006. https://www.ojp.gov/pdffiles1/nij/grants/213674.pdf.

[34] E.F. Law, K.B. Morris, Evaluating firearm examiner conclusion variability using cartridge case reproductions, 5, J. Forensic Sci. 66 (2021) 1704–1720, https://doi.org/10.1111/1556-4029.14758.

[35] A.H. Dorfman, R. Valliant, A Re-analysis of Repeatability and Reproducibility in the Ames-USDOE-FBI Study, 2022 submitted for publication, http://arxiv.org/abs/2204.08889.

[36] S. J Day, D.G. Altman, Blinding in Clinical Trials and Other Studies, BMJ (2000) 321, https://doi.org/10.1136/bmj.321.7259.504.

[37] T.A. Lang, D.F. Stroup, Who knew? The misleading specificity of 'double-blind' and what to do about it, Trials 21 (5 Aug.2020) 1. Gale Academic OneFile, link.gale.com/apps/doc/A631882619/AONE?u=anoñc2153463&sid=googleScholar&xid=4c421821. (Accessed 17 March 2022).

[38] I.E. Dror, Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias, Analytical Chemistry, 2020. https://pubs.acs.org/doi/pdf/10.1021/acs.analchem.0c00704.

[39] J.J. Koehler, Proficiency tests to estimate error rates in the forensic sciences, Law, Probability and Risk. https://academic.oup.com/lpr/article/12/1/89/924673, 2013, 12, 89-98.

[40] M.L. Pierce, L.J. Cook, Development and implementation of an effective blind proficiency testing program, Journal of Forensic Sciences 65 (3) (2020) 809–814, https://doi.org/10.1111/1556-4029.14269.

[41] 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 (May 2020) 3, https://doi.org/10.1111/1556-4029.14259.

[42] M. Neuman, C. Hundl, A. Grimaldi, D. Eudaley, D. Stein, P. Stout, Blind testing in firearms: Preliminary results from a blind quality control program, 00, J Forensic Sci. (2022) 1–11, https://doi.org/10.1111/1556-4029.15031.

[43] National Commission on Forensic Science, Ensuring that forensic analysis is based upon task-relevant information, 2015. https://www.justice.gov/archives/ncfs/file/818196/download. (Accessed 17 March 2022).

[44] I.E. Dror, M.L. Pierce, ISO standards addressing issues of bias and impartiality in forensic work, J Forensic Sci (2019), https://doi.org/10.1111/1556-4029.14265.

[45] R.J.A. Little, D.B. Rubin Statistical, Analysis with Missing Data, John Wiley & Sons, Inc., New York, 2002.

[46] R. Valliant, J.A. Dever, F. Kreuter, Practical Tools for Designing and Weighting Survey Samples, second ed., Springer, New York, 2018.