Articles

Bowling with Bumper Rails: How Firearms Examiners Have Duped the Courts and Generated Low Error Rates Only by Avoiding Challenging Comparisons

RICHARD E. GUTIERREZ[†]

[†] Assistant Professor of Law, University of Illinois Chicago (UIC) School of Law. I am grateful to Dean David Faigman and Dr. Nicholas Scurich for inviting me to speak at and submit this Article for UC San Fransisco's "Forensic Identification in Criminal Courts" symposium as well as to Margaret Armalas and Emily Prokesch for their invaluable insights and feedback about the piece (not to mention their incomparable commitment to defending the indigent-accused). Finally, this Article draws heavily from briefing I undertook when employed by the Law Office of the Cook County Public Defender in admissibility litigation concerning firearms examination evidence in the case of *Illinois v. Winfield*. I would therefore be remiss not to extend my thanks to that office for its foresight in establishing a unit specialized in forensic defense and as its unwavering support of litigation pushing back against the tide of firearms comparison testimony, as well to as my partners on that case: Margaret Domin, Ashley Shambley, Joseph Cavise, and Celeste Addyman.

TABLE OF CONTENTS

INTRODUCTION

Across its more than century-long history,[1] the field of firearms examination—a subspecies of forensic methodologies concerned with determining the gun that fired bullets or cartridge cases associated with a criminal offense[2]—has featured prominently in the aftermath of some of the United States' most infamous shootings. Practitioners have provided testimony in relation to, or otherwise "aided" with the investigation of, Sacco and Vanzetti, the St. Valentine's Day Massacre, and the assassination of President John F. Kennedy Jr.[3] But more recent years have seen cracks propagate wildly across the fragile veneer of science that allowed the field to infect our criminal legal system and insulate itself from meaningful scrutiny. We now know, in contrast to the handful of celebrated "successes" just noted, that misidentifications by firearms examiners have contributed to (if not single-handedly provoked) the wrongful arrest or conviction—for some nearly an execution—of at least seven men (Anthony Hinton, Leslie Merritt, Patrick Pursley, Desmond Ricks, Ricky Ross, Darrell Siggers, and Charles Stielow), have stolen more than a century's worth of freedom from the innocent, and have shuttered multiple forensic laboratories.[4] Though the field's defenders (and at times the courts) have too

1. *See* Brandon Garrett, Nicholas Scurich, Eric Tucker & Hannah Bloom, *Judging Firearms Evidence and the Rule 702 Amendments*, 107 JUDICATURE 41 (2023); James E. Hamby, *The History of Firearm and Toolmark Identification*, 31 AFTE J. 26 (1999).

2. *See* COMM. ON IDENTIFYING THE NEEDS OF THE FORENSIC SCIS. CMTY., NAT'L RSRCH. COUNCIL, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD 38, 150–51(2009) [hereinafter NAS REPORT]; PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS 23, 104 (2016) [hereinafter PCAST REPORT].

3. *See* Hamby, *supra* note 1.

4. *See* Craig Cooley & Gabriel Oberfield, *Symposium:* Daubert, *Innocence, and the Future of Forensic Science: Increasing Forensic Evidence's Reliability and Minimizing Wrongful Convictions: Applying* Daubert *Isn't the Only Problem*, 43 TULSA L. REV. 285, 337–38 (2007); Brandon L. Garrett, *Siggers' Firearms Exoneration*, DUKE L. FORENSIC F. (Oct. 23, 2018), https://sites.law.duke.edu/forensicsforum/2018/10/23/siggers-firearms-exoneration; *Siggers v. Alex*, No. 19-CV-12521, 2021 U.S. Dist. LEXIS 182956 (E.D. Mich. Sept. 24, 2021); Hinton v. Alabama, 571 U.S. 263 (2014); Daniella Silva, *Anthony Ray Hinton, Alabama Man Who Spent 30 Years on Death Row, Has Case Dismissed*, NBC NEWS (Apr. 2, 2015), https://www.nbcnews.com/storyline/lethal-injection/anthony-ray-hinton-alabama-man-who-spent-30-years-death-n334881; Ricks v. Pauch, No. 17-12784, 2020 U.S. Dist. LEXIS 50109 (E.D. Mich. Mar. 23, 2020); People v. Pursley, 2018 IL App (2d) 170227-U; Ivan Moreno, *Rockford Man Who Spent 23 Years in Prison Acquitted After Ballistics Retest Proves Innocence*, CHI. TRIB. (Jan. 16, 2019), https://www.chicagotribune.com/nation-world/ct-rockford-man-freed-after-ballistics-retest--20190116-story.html; Merritt v. Arizona, 425 F. Supp. 3d 1201 (D. Ariz. 2019); MICH. ST. POLICE FORENSIC SCI. DIV., AUDIT OF THE DETROIT POLICE DEPARTMENT FORENSIC SERVICES LABORATORY FIREARMS UNIT (2008), www.sado.org/content/pub/10559_MSP-DCL-Audit.pdf; Nick Bunkley, *Detroit Police Lab Is Closed After Audit Finds Serious Errors in Many Cases*, N.Y. TIMES (Sept. 25, 2008), https://www.nytimes.com/2008/09/26/us/26detroit.html; Erin Palmer, *The DC Crime Lab Symbolizes a Decade of Failures*, MEDIUM (Mar. 3, 2022), https://erinfordc.medium.com/the-dc-crime-lab-symbolizes-a-decade-of-failures-6cac4e1db791; SNA INT'L, DC DEPARTMENT OF FORENSIC SCIENCES LABORATORY ASSESSMENT REPORT (Dec. 8, 2021), https://dfs.dc.gov/sites/default/files/dc/sites/dfs/publication/attachments/DFS%20Forensic%20Laboratory%20Assessment%20Report.pdf.

often turned a blind eye to such injustice,[5] the same can no longer be said of independent scientists, who (at least in the last two decades) have taken note of the questionable practices and foundations of firearms examination and issued scathing critiques of its purported "standards," its underlying research, and even the method preferred by its practitioners for calculating error rates.[6] In turn, litigants have mounted sweeping challenges[7] to the field's methods under both *Daubert* and *Frye* admissibility standards,[8] forcing a growing trend in United States courts towards stringent limitations, if not outright exclusion, of testimony purporting to identify the source of fired bullets or cartridge cases.[9] Indeed, in contrast to the nearly uniform judicial record admitting firearms examination evidence just a few years ago,[10] at least eight judges have now ruled that the field's methods lack general acceptance, at least six have precluded testimony purporting to opine on the source of fired bullets and cartridge cases (with two others limiting source attribution testimony to whether or not a specific gun could be eliminated/excluded as the source of fired munition), and one has barred firearms examination testimony (even about the general markings and features of fired bullets and cartridge cases) outright.[11]

---

5. *See, e.g.*, Keith L. Monson, Erich D. Smith & Eugene M. Peters, *Authors' Response to Gutierrez et al. Commentary on Monson KL, Smith ED, Peters EM. Accuracy of Comparison Decisions by Forensic Firearms Examiners*, 68 J. FORENSIC SCIS. 1102, 1103 (2023) (quoting United States v. Diaz, No. CR 05-00167 WHA, 2007 WL 485967, at *13 (N.D. Cal. Feb. 12, 2007) ("[T]here has never been a single documented decision in the United States where an incorrect firearms identification was used to convict a defendant.")).

6. *See infra* Part II; NAS REPORT, *supra* note 2, at 150–56; PCAST REPORT, *supra* note 2, at 104–14.

7. *See, e.g.*, Illinois v. Winfield, No. 15CR14066-01 (Cir. Ct. Cook Cnty. Feb. 8, 2023) (on file with author); New York v. Ross, Ind. No. 267/2018, 129 N.Y.S.3d 629 (N.Y. Sup. Ct. Jan 23, 2020); United States v. Tibbs, No. 2016-CF1-19431, 2019 WL 4359486 (D.C. Super. Ct. Sep. 5, 2019).

8. *See* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 593–95 (1993) (noting that admissibility of expert testimony under Federal Rule of Evidence 702 "entails a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue" involving a "flexible" inquiry into the method's testability, whether the method has been subjected to peer review and publication, the method's potential error rate, the existence of standards governing the method, and whether the method has achieved general acceptance in the relevant scientific community); *see also* Nelson v. Tennessee Gas Pipeline Co., 243 F.3d 244, 251 (6th Cir. 2001) (explaining that "the factors mentioned in *Daubert* were neither definitive, nor exhaustive, and may or may not be pertinent to the assessment in any particular case") (citing Kumho Tire Co. v. Carmichael, 526 U.S. 137, 150 (1999)); Frye v. United States, 293 F. 1013 (D.C. Cir. 1923) ("[T]he thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs."). Federal courts, and those of most States, follow the *Daubert* standard, but a minority continue to employ *Frye*. *See, e.g.*, Andrew R. Stofli, *Note: Why Illinois Should Abandon Frye's General Acceptance Standard for the Admission of Novel Scientific Evidence*, 78 CHI.-KENT L. REV. 861, 862 (2003).

9. *See infra* Part II.

10. *See, e.g.*, United States v. Green, 405 F. Supp. 2d 104, 108, 123 (D. Mass. 2005) (noting the "serious deficiencies" of firearms examination and yet nevertheless concluding that admission was "compelled" because "every single court post-*Daubert*" had done so) (internal quotations omitted); Illinois vs. Rodriguez, 2018 IL App. (1st) 141379-B, at ¶ 59 (relying on the lack "of any published opinion of any court stating that firearms evidence was not generally accepted in the scientific community" to admit evidence from the field).

11. *See Winfield*, No. 15CR14066-01, at 32, 37, 41 (total exclusion and field lacks general acceptance); *Ross*, Ind. No. 267/2018, 129 N.Y.S.3d, at 642 (field lacks general acceptance and precluding opinion testimony regarding source); New York v. Terrell Lewis, Ind#1717/2020 (N.Y. Sup. Ct. Feb. 6, 2023) (following *Ross*)

But despite such burgeoning engagement with the substantial scientific shortcomings of firearms examination methods—engagement that has resulted in division amongst the courts regarding the field's ability to satisfy each and every one of *Daubert*'s testability, peer review, general acceptance, controlling standards, and error rate factors—judges remain curiously uniform (with just a few notable exceptions) in their acceptance of practitioner and law enforcement claims to the effect that the method's false positive (i.e. misidentification) rate hovers at or below 2 percent based on the results of repeated accuracy studies.[12] In other words, while courts have fallen out over whether a false positive rate of 2 percent weighs in favor of, or against, the admissibility of firearms examination evidence, they largely have not disputed the veracity of that figure,[13] and thus have necessarily ignored that scientists from outside the relatively insular community of firearms examination practitioners consistently balk at such estimates and the calculations used to derive them (citing, among other issues, the declared nature of existing accuracy studies, their inadequate sampling of participants, high rates of participant attrition, and *de minimus* data transparency).[14] But here's the all-the-more-troubling rub: Even following in the misguided footsteps of these courts, utilizing calculations preferred by practitioners, and dismissing scientific concerns about the trustworthiness of false positive estimates, accuracy studies of firearms examination methods have not uniformly documented misidentification rates of 2 percent or less. In fact, to date, four studies have produced false positive estimates indicating that examiners may misidentify the gun that fired a bullet or cartridge case in as many as 13.1 percent, 14.3 percent, 21.1 percent, or 39.6 percent of cases in which

(on file with author); United States v. Briscoe, No. 20-CR-1777 MV, 2023 WL 8096886, at *11–12 (D.N.M. Nov. 21, 2023) (field lacks general acceptance and precluding opinion testimony regarding source); Oregon v. Moore, No. 18CR77176, at 26, 29 (Cir. Ct. Or. Aug. 8, 2023) (field lacks general acceptance and precluding opinion testimony regarding source) (on file with author); United States v. Shipp, 422 F. Supp. 3d 762, 782–83 (E.D.N.Y. 2019) (field lacks general acceptance); United States v. Adams, 444 F. Supp. 3d 1248, 1266–67 (D. Or. Mar. 16, 2020) (field lacks general acceptance and precluding opinion testimony regarding source); United States v. Tibbs, No. 2016-CF1-19431, 2019 WL 4359486, at *21–22 (D.C. Super. Sep. 5, 2019) (field lacks general acceptance and restricting testimony to cannot exclude language); Missouri v. Goodwin-Bey, No. 1531-CR00555-01 (Cir. Ct. Green Cnty. Dec. 16, 2016) (restricting testimony to cannot exclude language) (on file with author). This list does not even include those decisions which have excluded outright testimony from firearms examiners on more case specific grounds. *See, e.g.*, Nebraska v. Kuek, No. Cr 19-2918 (Dist. Ct. Dec. 4, 2023) (reliance exclusively on magazine marks).

12. *See infra* Part II.

13. *Compare* United States v. Harris, 502 F. Supp. 3d 28, 39 (D.D.C. 2020) (citing false positive rates up to 1.6 percent before concluding: "Because the evidence shows that error rates for false identifications made by trained examiners is low—even under the PCAST's black-box study requirements—this factor also weighs in favor of admitting Mr. Monturo's expert testimony") *with Shipp*, 422 F. Supp. 3d at 778 ("[The s]tudy that most closely resembles fieldwork estimated that a firearms toolmark examiner may incorrectly conclude that a recovered piece of ballistics evidence matches a test fire once out of every 46 examinations. When compared to the error rates of other branches of forensic science—as rare as 1 in 10 billion for single source or simple mixture DNA comparisons—this error rate cautions against the reliability of the AFTE Theory.").

14. *See infra* Part II.

they evaluate different-source comparisons and reach conclusive source determinations.[15]

This Article takes as its focus that glaring asymmetry between the much-cited fantasy, and the nullifying reality, of false positive estimates for the field of firearms examination. Rather than retread ground eloquently and fulsomely explored elsewhere by scholars regarding the multitude of reasons for skepticism about the potential for misidentification of fired bullets and cartridge cases, it instead engages with judicial views of false positive rates where they lay, by (1) arguing that firearms examiners have achieved those oft-accepted misidentification figures of 2 percent or less only by stacking the deck in accuracy studies with simplistic comparisons, and (2) explaining why the studies that have documented far more prevalent occurrences of error better explore (critical to any validation effort) "the full range and distribution of types and difficulty normally seen in casework."[16] To that end, this Article begins with context: providing background on firearms examination methods in Part I, discussing calls for greater scrutiny of the empirical foundations of the field and court reactions thereto in Part II, and emphasizing the importance of including challenging comparisons in accuracy studies in Part III. Building from this substructure, Part IV then seeks a reckoning, with the failure of many firearms studies to test examiners robustly, and with the harrowing implications of those precious few that have bucked that trend and explored the true limits of the field's validity. If judges are to actuate their essential gatekeeping function—all the more in light of proposed amendments to Federal Rule of Evidence 702 which emphasize this responsibility and respond to the "incorrect application" of its mandates exemplified by treating "critical questions of the sufficiency of an expert's basis, and the application of the expert's methodology" as "questions of weight and not admissibility"[17]—then the time is ripe (indeed, long past due) for such an intervention into the misrepresentation of error rates that has so

15. *See* Alan Dorfman & Richard Valiant, *Inconclusives, Errors, and Error Rates in Forensic Firearms Analysis: Three statistical perspectives*, 5 FORENSIC SCI. INT'L: SYNERGY 100273 (2022) (citing Julie Knapp & Angela Garvin, *Consecutively Manufactured .25 Auto F.I.E. Barrels- A Validation Study*, Presentation at AFTE 43rd Annual Training Seminar (2012) (on file with author); Petra Pauw-Vugts, A. Walters, L. Øren & L. Pfoser, *FAID: Proficiency Test & Workshop*, 45 AFTE J. 115 (2013); Brandon A. Best & Elizabeth A. Gardner, *An Assessment of the Foundational Validity of Firearms Identification Using Ten Consecutively Button-Rifled Barrels*, 54 AFTE J. 28 (2022); Erwin J.A.T. Mattijssen, Cilia L. M. Witteman, Charles E. H. Berger, Nicolaas W. Brand & Reinoud D. Stoel, *Validity and Reliability of Forensic Firearm Examiners*, 307 FORENSIC SCI. INT'L, Feb. 2020. For further explanation of the calculations behind these figures and the nuances of the studies which produced them, see *infra* Part IV.

16. HUM. FACTORS TASK GRP., ORGANIZATION OF SCIENTIFIC AREA COMMITTEES FOR FORENSIC SCIENCE, HUMAN FACTORS IN VALIDATION AND PERFORMANCE TESTING OF FORENSIC SCIENCE, OSAC TECHNICAL SERIES 0004, at 11 (2020), https://www.nist.gov/system/files/documents/2023/10/26/OSACTechSeriesPub_HF%20in%20Validation%20and%20Performance%20Testing%20of%20Forensic%20Science_March2020.pdf [hereinafter OSAC HUMAN FACTORS REPORT]; *see generally infra* Part III.

17. Memorandum from John D. Bates, Chair, Committee on Rules of Practice and Procedure to Scott S. Harris, Clerk, Supreme Court of the United States, at 227 (Oct. 19, 2022), https://www.uscourts.gov/sites/default/files/2022_scotus_package_0.pdf.

swayed courts. Ultimately, no method with such flagrant potential to misidentify evidence deserves an opportunity to feature in determinations of guilt and thereby imperil the innocent. The age of the firearms examiner as expert witness must end.

I. IT'S SUBJECTIVE! IT'S CIRCULAR! IT'S FIREARMS EXAMINATION!!!

Firearms examination is a branch on the larger tree of forensic pattern-matching methods (and a sub-discipline within toolmark comparison more generally).[18] On the whole, such methods—also known as "feature comparison methods," and encompassing a wide range of disciplines developed to compare fingerprints, bitemarks, handwriting, and the like—"aim to determine whether an evidentiary sample (*e.g.*, from a crime scene) is or is not associated with a potential source sample (*e.g.*, from a suspect) based on the presence of similar patterns, impressions, features, or characteristics in the sample and the source."[19] Firearms examination in particular concerns itself with offering conclusions about the source of spent bullets and cartridge cases, in other words, "examiners attempt to determine whether ammunition is or is not associated with a *specific* firearm based on toolmarks produced by guns on the ammunition."[20] The idea, at its most basic level, is that when the harder metals of various gun components—the barrel interior, the parts of the chamber (including a firing pin, breech face, firing pin aperture, extractor, and ejector), and the magazine—come into contact with the softer metals of bullets and cartridges during the high pressure, high velocity, explosive firing process, they may scratch (striate) or stamp (impress) markings onto the latter's surfaces.[21]

The first step in the methodology of firearms examination involves an evaluation for "class characteristics," defined as "[m]easurable features of a specimen which indicate a restricted group source."[22] Essentially, such characteristics are those predetermined by the manufacturer of a firearm—such as the caliber, shape of the firing pin, or the number and twist of the lands and grooves within a barrel—and thus common to all guns of the same make and model.[23] At this stage, firearms examiners perform a "classification," as opposed to "individualization," function. They group guns into those that could, versus could not, have fired a particular bullet or cartridge case, but (because a great

18. *See* NAS REPORT, *supra* note 2, at 38, 150–51 (firearms examination simply involves comparison of the highly specialized toolmarks specific to the manufacture and use of guns, as opposed to marks left behind by other tools like screwdrivers, wire cutters, and the like).

19. *See* PCAST REPORT, *supra* note 2, at 23.

20. *Id.* at 104 (emphasis original).

21. *See generally* Robert M. Thompson, *Firearm Identification in the Forensic Laboratory*, NAT'L DIST. ATTY. ASS'N (2010), https://www.ojp.gov/ncjrs/virtual-library/abstracts/firearm-identification-forensic-science-laboratory.

22. ASS'N OF FIREARMS & TOOLMARK EXAM'RS, GLOSSARY 38 (6th ed. 2013), https://forensicresources.org/wp-content/uploads/2021/07/AFTE-Glossary-06-25-2021.pdf [hereinafter GLOSSARY].

23. *See, e.g.*, NAS REPORT, *supra* note 2, at 152.

many guns share the same class characteristics) they cannot single out any specific firearm as the source of spent ammunition.[24] In other words, they may reach an "elimination" conclusion if two items display different class characteristics (they may determine that a particular gun could not possibly have fired a particular bullet or cartridge case), but they must otherwise continue their examination.[25] Such "group-level" conclusions have not sparked much controversy,[26] although they can, in certain circumstances, become problematic.[27]

But, as noted above, firearms examiners do go beyond simply categorizing or classifying firearms, they seek to determine whether a *specific* firearm did or did not fire submitted bullets and cartridge cases. To reach such source determinations, firearms examiners compare the microscopic markings (scratches and impressions) left behind on fired bullets and cartridge cases.[28] Called "individual characteristics," these markings are "produced by the random imperfections or irregularities of tool surfaces," and result either from imperfection in the manufacturing process, or wear and tear on a firearm;[29] examiners must also distinguish these microscopic markings from so-called "subclass characteristics" (much more on these later)[30] which are "features that may be produced during manufacture that are consistent among items fabricated by the same tool in the same approximate state of wear," or, put another way, highly similar markings left behind during manufacturing on components produced consecutively.[31] The field relies on the comparison of so-called "individual characteristics" due to a belief that they are highly discriminating, if

24. *See, e.g.*, *id.* at 117–18.

25. *See, e.g.*, PCAST REPORT, *supra* note 2, at 104.

26. *See, e.g.*, NAS REPORT, *supra* note 2, at 8 (noting that "identification of a specific individual, they may still provide useful and accurate information about questions of classification."); Nicholas Scurich, David L. Faigman & Thomas D. Albright, *Scientific Guidelines for Evaluating the Validity of Forensic Feature-Comparison Methods*, 120 PROC. NAT'L ACADEMY SCIS. 2301843120, at 5 (2023) ("[E]xaminers ought to be limited to making general group- level statements, not individualized statements. An example of such testimony might be 'the bullet that killed the victim is consistent with having been shot from a .38 caliber Smith & Wesson, and there are approximately 10,000 such guns in circulation in the Southwest United States. Any one of those 10,000 guns could have left similar striae found on the *bullet*.'"); Ricks v. Pauch, No. 17-12784, 2020 U.S. Dist. LEXIS 50109, at *27 n.2 (E.D. Mich. Mar. 23, 2020) (calling class determinations "objective"); *New York v. Ross*, Ind. No. 267/2018, 129 N.Y.S.3d 629, at 641 (N.Y. Sup. Ct. Jan 23, 2020) ("It would be farcical to preclude experienced ballistics experts from rendering any opinion about known manufacturing marks. There is a consensus, or at least not all that much disagreement, to allow examiners to express an opinion on toolmarks that are class characteristics.").

27. *See, e.g.*, Gil Hocherman & Pavel Giverts, *Identification of Polygonal Barrel Sub-Family Characteristics*, 35 AFTE J. 197, 200 (2003) (describing examiner struggles to accurately categorize the class characteristics of polygonally-rifled barrels).

28. *See, e.g.*, PCAST REPORT, *supra* note 2, at 104.

29. GLOSSARY, *supra* note 22, at 65.

30. *See infra* Part IV.A.

31. GLOSSARY, *supra* note 22, at 118; *see* Alfred Biasotti & John Murdock, *Criteria for Identification or State of the Art of Firearm & Toolmark Identification*, 16 AFTE J. 16, 17 (1984); *see also* Adina Schwartz, *A Systemic Challenge to the Reliability & Admissibility of Firearms & Toolmark Identification*, 6 COLUM. SCI. & TECH. L. REV. 2 (2005).

not "unique"—that they will, in other words, appear highly similar on ammunition fired by the same gun and very distinct on ammunition fired by different guns.[32] But both versions of that internal dogma have been met by skepticism, with scholars so rebuking claims of uniqueness[33] that even law enforcement groups have more recently abandoned the term.[34] Lacking concrete, empirical assessments of the rarity or discriminability of particular arrangements of "individual characteristics," as well as "realistically large and complex databases" of known samples from which to develop them,[35] the comparison of individual characteristics remains dependent (as it has been since the field's inception over a century ago) on the subjective judgment of examiners.[36] Eventually, if early signs hold true,[37] comparison algorithms driven by 3D

32. *See* Thompson, *supra* note 21, at 16–25; Ass'n Firearms & Toolmark Exam'rs, Comm. for the Advancement of the Sci. of Firearm & Toolmark Identification, *Theory of Identification as It Relates to Toolmarks: Revisited*, 43 AFTE J. 287 (2011) ("The theory of identification as it pertains to the comparison of toolmarks enables opinions of common origin to be made when the unique surface contours of two toolmarks are in 'sufficient agreement.'") [hereinafter *Theory of Identification: Revisited*]; GLOSSARY, *supra* note 22, at 65 (calling individual characteristics "unique" to a particular tool).

33. *See, e.g.*, PCAST REPORT, *supra* note 2, at 62 ("The issue is not whether objects or features differ; they surely do if one looks at a fine enough level. The issue is how well and under what circumstances examiners applying a given metrological method can reliably detect relevant differences in features to reliably identify whether they share a common source"); Mark Page, Jane Taylor & Matt Blenkin, *Uniqueness in the Forensic Identification Sciences—Fact or Fiction?*, 206 FORENSIC SCI. INT'L 12, 14–15 (2011) ("Regardless of the method used to arrive at the probability of a particular forensic trait existing, extrapolation to uniqueness from these results still involves a 'leap of faith' . . . . The concept of 'uniqueness' has more the qualities of a cultural meme than a scientific fact."); NAT'L RSCH. COUNCIL, BALLISTIC IMAGING 3, 82 (Daniel L. Cork et al. eds., 2008) ("A significant amount of research would be needed to scientifically determine the degree to which firearms-related toolmarks are unique or even to quantitatively characterize the probability of uniqueness . . . . [T]he validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated.").

34. U.S. Dep't of Just., *Uniform Language for Testimony and Reports for the Firearms/Toolmark Discipline Pattern Examination* 3 (2020), https://www.justice.gov/d9/2024-01/final_firearms_pattern_examination_ultr_revision_effective_8.15.20.pdf.

35. *See, e.g.*, PCAST REPORT, *supra* note 2, at 114; Eric Hare, Heike Hofmann & Alicia Carriquiry, *Automatic Matching of Bullet Land Impressions*, 11 ANNALS APPLIED STAT. 2332, 2354 (2017) ("To understand whether an automated approach along the lines of the one we propose can accurately identify sets of bullets with undistinguishable markings, it will be necessary to assemble a much larger database that includes a wide range of ammunition types, degrees of damage, gun makes, etc. We are unaware of the existence of any such database. In addition to serving as a realistic testbed for the performance of the auto-mated matching, such a database would also permit testing the underlying, as of yet untested, assumptions of uniqueness and reproducibility of the markings left by a gun on bullets.").

36. *See Theory of Identification: Revisited*, supra note 32, at 287 ("Currently the interpretation of individualization/identification is subjective in nature, founded on scientific principles and based on the examiner's training and experience."). The use by some examiners of numerical criteria in the form of consecutive matching striae (CMS) does not alleviate the inherent subjectivity of bullet and cartridge case comparisons. *See* Stephen G. Bunch, *Consecutive Matching Striation Criteria: A General Critique*, 45 J. FORENSIC SCIS. 955, 959 (2000); *see also* Jerry Miller, *Criteria for Identification of Toolmarks Part III: Supporting the Conclusion*, 36 AFTE J. 7, 9 (2004) (documenting substantial variation in the line counting used in a cms approach).

37. *See, e.g.*, Fabiano Riva, Rob Hermsen, Erwin Mattijssen, Pascal Pieper & Christophe Champod, *Objective Evaluation of Subclass Characteristics on Breech Face Marks*, 62 J. FORENSIC SCIS. 417 (2017); John Song, Theodore V. Vorburger, Wei Chu, James Yen, Johannes A. Soons, Daniel B. Ott & Nien Fan Zhang,

topography scanning and complex statistics will "convert firearms analysis from a subjective method to an objective method."[38] But despite the promise they have shown (including by outperforming human examiners at distinguishing bullets fired by different guns),[39] "more work must be conducted before they can be implemented in real case work."[40]

Until such time, examiners ply their trade, following guidance by their field's lead professional association, the Association of Firearm and Tool Mark Examiners (AFTE),[41] in a document called the *Theory of Identification as it Relates to Toolmarks*, by placing two bullets or cartridge cases under a light comparison microscope (two separate microscopes connected by an optical bridge) and looking for the presence or absence of what that group calls "sufficient agreement," defined (without numerical thresholds or guideposts of any kind) as a level of agreement that "exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool."[42] In other words, a match is a match if it looks like other matches an examiner has seen, and looks more similar than non-matches the examiner has seen (a reality that all but guarantees that "there will be some difference between examiners as to what constitutes the best-known non-match situation").[43]

Scientists outside the field of firearms examination have recoiled from the *Theory of Identification* and its description of "sufficient agreement," faulting the discipline for "unarticulated standards," noting that "a fundamental problem with toolmark and firearm analysis is the lack of a precisely defined process," criticizing the *Theory of Identification* for failing to "provide a specific protocol," as well as "not even consider[ing], much less address[ing], questions regarding variability, reliability, repeatability, or the number of correlations needed to achieve a given degree of confidence,"[44] outright dismissing it as "circular,"[45] and even opining that it "contemplates memory and analytical

Estimating Error Rates for Firearm Evidence Identifications in Forensic Science, 284 FORENSIC SCI. INT'L 15 (2018)

38. PCAST REPORT, *supra* note 2, at 113.

39. *See, e.g.*, Melissa Nally, *Ruger LCP Study: A two-pronged approach*, CTR. FOR STAT. & APP. IN FORENSIC EVIDENCE (2021), https://dr.lib.iastate.edu/entities/publication/5fc21111-3f89-4684-9743-23e00eb8713b.

40. Heike Hofmann, Alicia Carriquiry & Susan Vanderplas, *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 L., PROBABILITY & RISK 317, 344 (2020).

41. *See* Ass'n Firearms & Toolmark Exam'rs, Comm. for the Advancement of the Sci. of Firearm & Toolmark Identification, *The Response of the Association of Firearms & Tool Mark Examiners to the National Academy of Sciences 2008 Report Assessing the Feasibility, Accuracy, & Technical Capacity of a National Ballistics Database*, 40 AFTE J. 234, 237 (2008).

42. *Theory of Identification: Revisited*, supra note 32, at 287; *see* Thompson, *supra* note 21, at 8–12.

43. Ronald G. Nichols, *The Scientific Foundations of the Firearms & Toolmark Identification: Responding to Recent Challenges*, CAC NEWS, at 26 (2nd Quarter 2006), http://www.forensicdna.com/assets/2ndq06.pdf.

44. NAS REPORT, *supra* note 2, at 153–55.

45. PCAST REPORT, *supra* note 2, at 60, 104.

capacities that are implausible."[46] But none of that has stopped firearms examiners from adopting confident, categorical, and certain conclusions to go along with their deeply-subjective and rudderless approach. Specifically, and again pursuant to AFTE guidance, at the end of a given comparison,[47] examiners may reach one of three conclusions: (1) identification (the bullets or cartridge cases were fired from the same gun); (2) elimination (the fired bullets or cartridge cases were fired by different guns); or (3) inconclusive (the individual characteristics do not display sufficient agreement for an identification or sufficient disagreement for an elimination).[48] Reporting practices vary by laboratory, with some splitting the inconclusive category into three (to reflect separate bins for cases involving some agreement or some disagreement of individual characteristics), and most either refusing outright to reach eliminations based on individual characteristics, regardless of the extent of disagreement observed, or doing so only in exceptional circumstances.[49] But in stark contrast to widespread hesitancy to eliminate, examiners show little restraint when it comes to producing powerful incriminating evidence: when an examiner renders an "identification" conclusion, they may opine that the potential for error "is so remote as to be considered a practical impossibility."[50] Though, as this Article will demonstrate, such grandiose claims fall far afield of

46. Scurich et al., *supra* note 26, at 4.

47. As discussed in more detail in Part IV, most labs follow up an initial examiner's conclusion with a quality assurance step called verification in which a second examiner checks the work of the first. *See* Sci. Working Grp. for Firearms & Toolmarks (SWGGUN), *Systemic Requirements / Recommendations for the Forensic Firearm & Toolmark Laboratory*, at 4 (2016), https://www.nist.gov/organization-scientific-area-committees-forensic-science/firearms-toolmarks-subcommittee. How exactly labs accomplish this, however, varies wildly. *See, e.g.*, Best & Gardner, *supra* note 15, at 35.

48. *See Range of Conclusions*, ASS'N OF FIREARM & TOOLMARK EXAM'RS, https://afte.org/about-us/what-is-afte/afte-range-of-conclusions (last visited June 6, 2024).

49. *See* Maneka Sinha & Richard E. Gutierrez, *Signal Detection Theory Fails to Account for Real-World Consequences of Inconclusive Decisions*, 21 L., PROBABILITY & RISK 131, 132 (2022); Best & Gardner, *supra* note 15, at 36; Sci. Working Grp. for Firearms & Toolmarks (SWGGUN), *Elimination Factors Related to FA/TM Examinations*, at 1 (2016), www.nist.gov/organization-scientific-area-committees-forensic-science/firearms-toolmarks-subcommittee; Keith L. Monson, Erich D. Smith & Stanley J. Bajic, *Planning, Design and Logistics of a Decision Analysis Study: The FBI/Ames Study Involving Forensic Firearms Examiners*, 4 FORENSIC SCI. INT'L: SYNERGY, 2022, at 6; Laura Knowles, Daniel Hockey & John Marshall, *The Validation of 3D Virtual Comparison Microscopy (VCM) in the Comparison of Expended Cartridge Cases*, 67 J. FORENSIC SCIS. 516, 522 (2021); David P. Baldwin, Stanley J. Bajic, Max D. Morris & Daniel S. Zamzow, *A Study of Examiner Accuracy in Cartridge Case Comparisons. Part 1: Examiner Error Rates*, 349 FORENSIC SCI INT'L, Aug. 2023, at 5, 7 (2023); David P. Baldwin, Stanley J. Bajic, Max D. Morris & Daniel S. Zamzow, *A Study of Examiner Accuracy in Cartridge Case Comparisons. Part 2: Examiner Use of the AFTE Range of Conclusions*, 349 FORENSIC SCI INT'L, Aug. 2023, at 3 (2023). The slant of firearms examination against eliminations was enough to shock even a seasoned reporter on forensic issues, who noted that "I learned something that after 20 years on this beat still managed to astonish me . . . . They refuse to exclude one specific gun if it would benefit the defense, but they're willing to exclude every gun in existence but one to benefit the prosecution. And this isn't the secret, unstated policy of a few rogue analysts. It's the official policy of some of the most widely-used and respected crime labs in the country." Radley Balko, *Devil in the Grooves: The Case Against Forensic Firearms Analysis*, WATCH (May 5, 2023), https://radleybalko.substack.com/p/devil-in-the-grooves-the-case-against.

50. *Theory of Identification: Revisited*, *supra* note 32, at 287.

reality,[51] the field of firearms examination has largely refused to dial them back.[52]

II. LET THE CRITICS HIT THE FLOOR: CALLS FOR A RIGOROUS EMPIRICAL FOUNDATION AND RESPONSES FROM THE COURTS

Although troubling in its completeness (other comparison methods in forensics at least utilize quasi-objective guideposts to moderate examiner decision making),[53] the subjectivity of firearms examination has not alone disqualified the field in the eyes of outside scientists.[54] Rather, the thrust of most criticisms has emphasized that methods deeply dependent upon human judgment, like firearms examination, necessitate "careful scrutiny . . . [because] they are especially vulnerable to human error, inconsistency across examiners, and cognitive bias."[55] In other words, since a human being "serves as an instrument for information measurement and classification[, and, a]s for any such instrument, we'd like to know how well it works," empirical studies of accuracy and consistency are necessary.[56] Far from controversial, such views

51. *See, e.g.*, PCAST REPORT, *supra* note 2, at 19 (describing as "scientifically indefensible" claims of: "a chance of error so remote as to be a 'practical impossibility'"); Simon A. Cole, "*Individualization is Dead, Long Live Individualization! Reforms of Reporting Practices for Fingerprint Analysis in the United States*, 13 L., PROBABILITY & RISK 117 (2014) (describing practical certainty as "an obscure and seemingly nonsensical value for a probability" and concluding that "neither the Theory of Identification nor the toolmark literature provides a defensible justification for claims that toolmark analysis can reduce the probability that two impressions derive from different sources to 'practical impossibility'"); William Tobin & Peter Blau, *Hypothesis Testing of the Critical Underlying Premise of Discernible Uniqueness in Firearms-Toolmark Forensic Practice*, 53 JURIMETRICS 121, 131 (2013) (calling on firearms examiners to "curb the excesses" of their conclusions and noting that "the switch to weaker forms of source attribution (such as 'practical certainty') is a cosmetic change that does nothing to remedy the underlying scientific shortcomings of F/TM practice").

52. *See, e.g.*, U.S. Dep't of Just., *Uniform Language for Testimony and Reports for the Firearms/Toolmark Discipline Pattern Analysis* (retaining term "source identification," refusing to prohibit expressions of practical certainty, and encouraging examiners describe "the probability that the two toolmarks were made by different sources" as "so small that it is negligible).

53. *See, e.g.*, Sci. Working Grp. on Friction Ridge Analysis, Stud. & Tech., *Document #10 Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint)*, (2013) (providing a sufficiency graph based on feature counts and quality to distinguish unidentifiable, complex, and non-complex latent prints); Org. of Sci. Area Comms Human Forensic Biology Subcommittee, *2021-S-0003 Standards for Determining Analytical and Stochastic Thresholds for Application to Forensic DNA Casework Using Electrophoresis Platforms* (2022) (discussing development and use of thresholds for determining when genetic markers may be distinguished from instrument noise or paired together), https://www.nist.gov/system/files/documents/2022/06/06/OSAC%202021-S-0003%20Standards%20for%20Determining%20Analytical%20and%20Stochastic%20Thresholds_OPEN%20COMMENT%20VERSION.pdf.

54. *Cf.* PCAST REPORT, *supra* note 2, at 105 ("[I]t is not necessary that toolmarks be unique for them to provide useful information whether a bullet may have been fired from a particular gun. However, it is essential that the accuracy of the method for comparing them be known based on empirical studies.").

55. *Id.* at 49. For definitions of terms critical to understanding metrics for empirical assessment of subjective methods see OSAC HUMAN FACTORS REPORT, *supra* note 16, at 3–6.

56. Thomas D. Albright, *How to Make Better Forensic Decisions*, 119 PROC. NAT'L ACADEMY SCIS. 2206567119, at 9 (2022); *see* PCAST REPORT, *supra* note 2, at 47, 49 ("Foundational validity for a forensic-science method requires that it be shown, based on empirical studies, to be repeatable, reproducible,

merely restate "the fundamental principles of the scientific method . . . that valid scientific knowledge can only be gained through empirical testing of specific propositions."[57] But their application to firearms examination did not occur until late in the life of that field, beginning with a trickle of articles critical of forensic methods in leading scientific journals,[58] continuing with a few pioneers specifically exploring the research base of firearms examination,[59] and ballooning with the publication of a trilogy of reports by scientific advisory bodies to the federal government variously concluding: (1) that "the validity of the fundamental assumptions . . . of firearms-related toolmarks has not yet been fully demonstrated;"[60] (2) that "no forensic method [other than DNA] has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source;"[61] and (3) that the research base of firearms examination "falls short of the scientific criteria for foundational validity."[62]

Perhaps because of its development, not in the traditional proving ground of research universities, but "heuristically" in crime labs and for the benefit of

---

and accurate, at levels that have been measured and are appropriate to the intended application. . . . Since the black box in the examiner's head cannot be examined directly for its foundational basis in science, the foundational validity of subjective methods can be established only through empirical studies of examiner's performance to determine whether they can provide accurate answers."); NAS REPORT, *supra* note 2, at 122 ("The assessment of the accuracy of the conclusions from forensic analyses and the estimation of relevant error rates are key components of the mission of forensic science"); Itiel E. Dror & Nicholas Scurich, *(Mis)use of Scientific Measurements in Forensic Science*, 2 FORENSIC SCI. INT'L: SYNERGY 333 (2020) ("One critical measurement metric in all sciences, and in forensic science in particular, are error rates, the topic of this article. Knowing the error rates in a particular forensic domain is a vital measurement needed to ascertain the weight of the evidence. The appropriate weight of the evidence cannot be known without some sense of the rates at which the technique errs."); Hofmann et al., *supra* note 40, at 318–19 (2020) (discussing the "need for scientific validation and experimentally determined error rates"); OSAC HUMAN FACTORS REPORT, *supra* note 16, at 6–8, 28 ("Why should forensic scientists conduct empirical studies to assess the accuracy of their methods? Validation is necessary in all scientific disciplines. It is particularly important in forensic science because of the consequences that may follow from a single forensic science analysis or comparison. The judgments of a DNA analyst, latent print examiner or tool mark examiner, based on a single comparison, can have dramatic consequences for human lives—a fact that the forensic science and legal communities know and acknowledge. The manifest importance of forensic science findings to the justice system makes it vital to have data on their accuracy.").

57. PCAST REPORT, *supra* note 2, at 46. Indeed, even groups including forensic practitioners have acknowledged as much. *See, e.g.*, Jonathan Koehler, Jennifer L. Mnookin, Simon A. Cole, Barry A.J. Fisher, Itiel E. Dror, Max Houck, Kieth Inman, David H. Kaye, Glenn Langenburg, D. Michel Risinger, Norah Rudin & Jay Siegel, *The Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 725, 745, 749 (2011).

58. *See, e.g.*, Donald Kennedy, *Forensic Science: Oxymoron?*, 302 SCIENCE 1625 (2003); Nature Editorial Bd., *Science in Court*, 464 NATURE 325 (2010); Michael J. Saks & Jonathan L. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 SCIENCE 892 (2005).

59. *See* Schwartz, *supra* note 31; Tobin & Blau, *supra* note 51; Clifford Spiegelman & William A. Tobin, *Analysis of Experiments in Forensic Firearms/Toolmark Practice Offered as Support for Low Rates of Practice Error & Claims of Inferential Certainty*, 12 L., PROBABILITY & RISK 115 (2013).

60. BALLISTIC IMAGING, *supra* note 33, at 82 (noting that significant additional research would be needed to place even the basic premises of firearms examination on "solid scientific footing").

61. NAS REPORT, *supra* note 2, at 7, 107–08 (emphasizing that firearms examination lacks "any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline").

62. PCAST REPORT, *supra* note 2, at 111.

law enforcement,[63] the field of firearms examination waited decades before producing the types of accuracy studies such critics saw as missing.[64] The first research effort consciously designed to estimate methodological error not released until the late 1990s.[65] Since then, researchers have conducted over a dozen such studies,[66] though their results have often, even at first blush, done more to spark concern than instill comfort. Fears about subjectivity begetting inconsistent and individualized examiner conclusion criteria have been vindicated by repeatability and reproducibility figures, with one study finding that examiners (looking at exactly the same bullets and cartridge cases) disagree with themselves 35.5 percent of the time, and with each other, a whopping 63.5 percent of the time.[67] And specificity rates (which measure the percentage of times examiners correctly eliminate on different source comparisons) have repeatedly fallen within a range approaching, or statistically worse, than chance

63. NAS REPORT, *supra* note 2, at 128; Maneka Sinha, *Radically Reimagining Forensic Evidence*, 73 ALA. L. REV. 879, 894–904 (2022) (describing the "carceral culture" of forensic methods, and concluding that as "[a]consequence of these law enforcement origins . . . forensic methods developed insulation from traditional scientific checks and balances like independent review, critique, and repeated testing, and in turn, a scientific culture designed to promote these features did not emerge").

64. Prior to conducting studies designed specifically to estimate error rates, the field of firearms examination and its practitioners did undergo proficiency testing. *See* J. L. Peterson & P. N. Markham, *Crime Laboratory Proficiency Testing Results, 1978-1991, I: Identification & Classification of Physical Evidence*, 40 J. FORENSIC SCIS. 994, 997 (1995). Indeed, some courts have relied on and discussed the results of such tests when confronting the issue of an error rate for the discipline. *See, e.g.*, United States v. Otero, 849 F. Supp. 2d 425, 433–34 (D.N.J. 2012); United States v. McCluskey, No. CR 10-2734 JCH, 2013 WL 12335325, at *7 (D.N.M. Feb. 7, 2013). But this Article does not linger on such tests because, even by the admission of firearms examination's proponents and major test developers, they present too simplistic a challenge (and are otherwise not designed) to serve as an appropriate estimate of error. *See* Collaborative Testing Services Inc., *CTS Statement on the Use of Proficiency Testing Data for Error Rate Determinations* (2010); PCAST REPORT, *supra* note 2, at 57–59; Adina Schwartz, *Challenging Firearms and Toolmark Identification - Part Two*, 32 CHAMPION 44, 47 (2008); Richard Grzybowski, Jerry Miller, Bruce Moran & John Murdock, *Firearm/Toolmark Identification: Passing the Reliability Test Under Federal and State Evidentiary Standards*, 35 AFTE J. 209, 219 (2003); Angela Stroman, *Empirically Determined Frequency of Error in Cartridge Case Examinations Using a Declared Double-Blind Format*, 46 AFTE J. 157, 158 (2014); Pauw-Vugts et al., *supra* note 15, at 117; Simon A. Cole, *More Than Zero: Accounting for Error in Latent Fingerprint Identifications*, 95 J. CRIM. L. & CRIMINOLOGY 985, 1029 (2005).

65. *See* David J. Brundage, *The Identification of Consecutively Rifled Gun Barrels*, 30 AFTE J. 438 (1998); Federal Bureau of Investigation, *Response to the Declaration Regarding Firearms and Toolmark Error Rates Filed in* Illinois v. Winfield, at 18 (May 3, 2022) (on file with author) (providing table of known error rates studies beginning, at the earliest with the Brundage study from 1998) [Hereinafter FBI Statement] ; United States Department of Justice, *Statement on the PCAST Report: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, at 23 (2021), https://www.justice.gov/opa/pr/justice-department-publishes-statement-2016-presidents-council-advisors-science-and (same) [Hereinafter DOJ Statement].

66. *See* FBI Statement, *supra* note 65, at 18–20; DOJ Statement, , *supra* note 65, at 23–24.

67. *See* Keith L. Monson, Erich D. Smith & Eugene M. Peters, *Repeatability and Reproducibility of Comparison Decisions by Firearms Examiners*, 68 J. FORENSIC SCIS. 1721 (2023); *see also* Alan H. Dorfman & Richard Valliant, *A Re-Analysis of Repeatability and Reproducibility in the Ames-USDOE-FBI Study*, 9 STAT. & PUB. POL'Y 175, 182 (2022) (concluding, after conducting widely accepted statistical analysis of said figures that they show "rather weak Repeatability and Reproducibility").

(as low as 13.1 percent),[68] leading critics to accuse the field of systemic bias against criminal defendants (who most commonly benefit from such conclusions).[69] Firearms examination has weathered these revelations largely due to the more narrow focus of its proponents, some outside scientists, and multiple courts on the false positive (*i.e.*, misidentification) rate,[70] which practitioners and law enforcement groups have consistently reported as below 2 percent.[71] But this myopic lens troublingly fails to account for the reality that "an erroneous individualization is only one of many ways in which [forensic] error can be implicated in wrongful convictions;" such methods can spell doom for criminal defendants, not just by wrongly implicating them, but also by failing

68. *See* Best & Gardner, *supra* note 15, at 32, 35 (23/176, 4 within class eliminations x 44 participants =176 within class eliminations possible); *accord* Jaimie A. Smith, *Beretta Barrel Fired Bullet Validation Study*, 66 J. FORENSIC SCIS. 547, 552 (2021) (22.8%); Baldwin et al., *supra* note 49, at 5 (65.24%); Max Guyll, Stephanie Madon, Yueran Yang, Kayla A. Burd & Gary Wells, *Validity of Forensic Cartridge-Case Comparisons*, 120 PROC. NAT'L ACADEMY SCIS. e2210428120, at 5 (2023) (63.5%); Keith L. Monson, Erich D. Smith & Eugene M. Peters, *Accuracy of Comparison Decisions by Forensic Firearms Examiners*, 68 J. FORENSIC SCIS. 86, 93 (2023) (33.8% for bullet comparisons and 48.5% for cartridge case comparisons).

69. *See* Sinha & Gutierrez, *supra* note 49, at 132–34 (tracking massive divides between sensitivity and specificity across multiple validation studies of firearms examination and concluding that "approach to inconclusive decisions fundamentally prejudices the accused, who is regularly deprived of exculpatory evidence by firearms examiners' inability or unwillingness to reach exclusion decisions, while inflicting no commensurate penalty on the prosecution"); Hofmann et al., *supra* note 40, at 342 ("[I]n the absence of definitive information, examiners tend to more often conclude identification than elimination . . . this results in a bias in favour of the prosecution."); Dorfman & Valiant, *supra* note 15, at 2 (2022) ("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."); Andrew M. Smith & Gary L. Wells, *Telling Us Less Than What They Know: Expert Inconclusive Reports Conceal Exculpatory Evidence in Forensic Cartridge-Case Comparisons*, 13 J. APPLIED RSCH. MEMORY & COGNITION 147, 152 (2023) (analyzing the field of firearms examination's "huge bias against reporting exculpatory evidence by hiding the exculpatory evidence in an inconclusive category"); Balko, *supra* note 49 (quoting Chris Fabricant of the Innocence Project as decrying that "When prosecutors send a bullet and gun to one of these labs for testing, they have nothing to lose. . . . At worst, they'll be told there's insufficient evidence to match the bullet to the gun, at which point they can just fall back on whatever other evidence they may have. But there's no risk of them hurting their case").

70. *See, e.g.*, PCAST REPORT, *supra* note 2, at 50 ("The false positive rate is especially important because false positive results can lead directly to wrongful convictions."); Raymond Valerio & Nelson Bunn, *Firearm Forensics Has Proven Reliable in the Courtroom. And in the Lab*, SCI. AM. (Nov. 27, 2023), https://www.scientificamerican.com/article/firearm-forensics-has-proven-reliable-in-the-courtroom-and-in-the-lab ("But inconclusive decisions do not send people to jail—identifications do…when judging reliability, the false positive error rate is paramount/"); United States v. Harris, 502 F. Supp. 3d 28, 39 (D.D.C. 2020) ("[T]he critical inquiry under this factor is the rate of error in which an examiner makes a false positive identification, as this is the type of error that could lead to a conviction premised on faulty evidence.").

71. *See, e.g.*, *Harris*, 502 F. Supp. 3d at 39 (noting that firearms examiner called by prosecution testified that "he had seen a rate of false positives in research studies ranging from 0–1.6 percent"); Valerio & Bunn, *supra* note 70 ("When an examiner opines that a fired casing came from a particular firearm, they are accurate more than 99 percent of the time."); Jim Agar, *The Admissibility of Firearms and ToolMarks Expert Testimony in the Shadow of PCAST*, 74 BAYLOR L. REV. 93, 193 (2022) (providing same figure); FBI Statement, *supra* note 65, at 3 ("[S]tudies produced low false positive rates of approximately 1% or less.").

to provide them with the concrete exculpatory evidence their innocence warrants.[72]

All the more problematically for the field, however, outside research scientists have almost uniformly expressed skepticism about false positive figures for firearms examination and the quality of the studies underlying them. Much of their criticism has focused on the decision of many study designers to utilize "set-based" approaches in which examiners compare multiple known and questioned bullets or cartridges simultaneously.[73] Despite the prevalence of such designs in the literature underlying firearms examination, scientists and mathematicians have pointed out that they (1) "ensur[e] that it is not possible to calculate the overall error rate, the correct decision rate, or the true negative rate (the specificity)," (2) "can inflate examiners' performance by allowing them to take advantage of internal dependencies in the data," and (3) even "have an inherent bias, because they . . . prevent evaluation of examiners on their ability to distinguish between different sources . . . so that in court they provide useful (but misleading) information to the prosecution while offering nothing useful to the defense."[74] But scholars have not limited themselves to attacking set-based

72. Simon A. Cole & Barry Sheck, *Fingerprints and Miscarriages of Justice: 'Other' Types of Error and A Post- Conviction Right to Database Searching*, 81 ALBANY L. REV. 807, 810, 819 (2018); *see* Abruquah v. Maryland, 296 A.3d 961, n.21 (2023) ("Although false positives create the greatest risk of leading directly to an erroneous guilty verdict, an examiner's erroneous failure to eliminate the possibility of a match could also contribute to an erroneous guilty verdict if the correct answer—elimination—would have led to an acquittal."); Sinha & Gutierrez, *supra* note 49, at 133 ("Missed exclusions may leave the innocent languishing in custody while investigators attempt to develop other evidence of guilt, bias investigators against pursuing other credible suspects, or even contribute to a wrongful conviction by depriving the accused of exculpatory evidence.").

73. *See* PCAST REPORT, *supra* note 2, at 106; OSAC HUMAN FACTORS REPORT, *supra* note 16, at 14.

74. Hofmann et al., *supra* note 40, at 332–33. These same views have been expressed by a multitude of scholars and have even been acknowledged as sound by firearms examiners themselves. *See, e.g.*, PCAST REPORT, *supra* note 2, at 106, 109 (disparaging the inherent data interdependencies of set-based studies, likening them to a game of Sudoku, and emphasizing that their potential for inflating the accuracy of examiners "is not just a theoretical possibility: it is evident in the results themselves. Specifically, the closed-set studies have inconclusive and false-positives rate that are dramatically lower (by more than 100-fold) than those for the partly open design (Miami-Dade study) or fully open, black-box designs (Ames Laboratory) studies described below (Table 2)"); Scurich et al., *supra* note 26, at 4 ("[W]hile these studies have been presented in court by FATM examiners as precisely the empirical support that science demands (31), these fundamental design flaws are now widely recognized as precluding their ability to measure a false positive error rate. Simply put, the studies did not measure what they claimed to measure, and consequently, their results have been misrepresented in court."); Baldwin et al., *supra* note 49, at 1 (noting that previous set-based studies "did not include truly independent sample sets that would allow the unbiased determination of false Identification or false Elimination error rates from the collected data"); Stroman, *supra* note 64, 160 ("[I]f the participants know they are dealing with a closed set of unknowns they will likely perform better on the test than if it were an open set because they may be able to use a process of elimination to infer at least a couple of the answers if they were able to identify the rest of the unknowns."); Guyll et al., *supra* note 68, at 2 ("[A] closed-set design only requires examiners to find the cartridge case that most closely matches another cartridge case to render a correct identification decision, a strategy that would be ineffective and inappropriate in the field. Thus, a closed-set design is ill-suited to establishing validity because it potentially overestimates accuracy and underestimates error."); Monson et al., *supra* note 49, at 5 ("[U]nderestimation of false positives [is] inherent in a closed set."); James E. Hamby, David J. Brundage, Nicholas D. K. Petraco & James W. Thorpe, *A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels—Analysis of Examiner Error Rate*, 64 J. FORENSIC SCIS. 551, 556 (2019) (describing criticisms of set-based studies as "appropriate").

studies as "inapposite for measuring examiner performance."[75] They have also pointed out a host of deficiencies even in more appropriately designed sample-to-sample (or pairwise) studies that eliminate interdependencies and allow for false positive rate calculations by "present[ing] test specimens as a series of pairs, asking the examiner to judge whether each pair of specimens has a common source, before presenting the next pair" (*i.e.*, one known and one questioned sample at a time).[76] Specifically, outside scientists and mathematicians have emphasized that existing studies of firearms examination likely underestimate the field's false positive rate because of issues ranging from a lack of control groups and inappropriate sampling of participants to unacceptable rates of participant attrition and declared rather than blind formats.[77]

Indeed, debate has swept up even the manner of calculating a misidentification rate given the prevalence of inconclusive conclusions used by examiners on different-source comparisons.[78] Examiners (so the argument goes) aware they are being tested, may default to saying inconclusive on difficult cases rather than reach an erroneous source conclusion, thereby "mask[ing] what would be a mistaken identification or elimination in casework" and "substantially reduc[ing] the credibility and reliability of the error rates reported."[79] While firearms examiners (and their allies) have preferred to

75. Scurich et al., *supra* note 26, at 4.

76. OSAC HUMAN FACTORS REPORT, *supra* note 16, at 14; PCAST REPORT, *supra* note 2, at 110.

77. *See, e.g.*, Scurich et al., *supra* note 26, at 4–6 (discussing a lack of control groups and declared nature of testing as well as sampling and attrition issues); Khori Khan & Alicia Carriquiry, *Shining a Light on Forensic Black-Box Studies*, 10 STAT. & PUB. POL'Y 2216748 (2023) (discussing sampling, attrition, and data transparency issues); Amicus Brief in Support of Appellant, Abruquah v. Maryland, COA-REG-0010-2022 (Ct. App. Sept. 2, 2022) (on file with author) (discussing problems with declared nature of tests, attrition, and sampling bias); Jonathan J. Koehler, *Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences*, 49 ARIZ. ST. L.J. 1369, 1409–14 (2017) (outlining similar flaws as applicable to studies of latent print examination studies); Hofmann et al., *supra* note 40, at 343 (discussing range of design flaws that undercut value of firearms examination validation studies). Indeed, as to the issue of declared testing and resulting changes to examiner performance, one scholar recently modeled the impact of even "small reductions in the threshold for identification, which might plausibly arise from an examiner's exposure to task-irrelevant information," (*i.e.*, information for investigators about the nature of a criminal case) and found that they "can dramatically increase the risk of convicting an innocent person," meaning that declared testing provides, at best, a deeply imperfect view of casework accuracy. William C. Thompson, *Shifting Decision Thresholds Can Undermine the Probative Value and Legal Utility of Forensic Pattern-Matching Evidence*, 120 PROC. NAT'L ACADEMY SCIS. e2301844120 (2023).

78. *See, e.g.*, Sinha & Gutierrez, *supra* note 49, at 132 (documenting inconclusive rates far great on different source as opposed to same source comparisons across multiple pairwise validation studies for firearms examination); Hofmann et al., *supra* note 40, at 344 ("[E]xaminers working under the AFTE range of conclusions appear to have a lower threshold for identification than for elimination; when evidence originates from different sources, examiners are more likely to arrive at an inconclusive decision than they are when the evidence has the same source.").

79. Dorfman & Valiant, *supra* note 15, at 7; *see also, e.g.*, Hal R. Arkes & Jonothan J. Koehler, *Inconclusives and Error Rates in Forensic Science: A Signal Detection Theory Approach*, 20 L., PROBABILITY & RISK 153, 165 (2021) ("If examiners do adopt different thresholds in test versus casework situations then the error rates identified from tests cannot be trusted as estimates of casework error rates even similar types of

calculate a false positive rate by including inconclusive results in the denominator,[80] and thus functionally treating them as correct[81] (false positive rate = false positives / (false positives + inconclusives + eliminations)), these concerns have led outside scholars to recommend either dropping inconclusive determinations from error rate calculations entirely as a "pass" (false positive rate = false positives / (false positives + eliminations)),[82] or—until studies incorporate metrics for gauging the propriety of inconclusive responses for specific comparisons—treating them as potential errors and including them in the numerator (false positive rate = (false positives + inconclusives) / (false positives + inconclusives + eliminations)).[83] Convincing grounds exist for following the last of those approaches—including that examiners, in contrast to AFTE's definition of inconclusive as reserved for "absence, insufficiency, or lack of reproducibility"[84] of individual characteristics, appear to deploy that conclusion even when comparing samples which display "extensive" markings[85]—but regardless of where one falls, it is difficult to dismiss the debate

samples and methods are involved. This is why it is important to consider how to implement blind proficiency testing throughout the forensic sciences."); Albright, *supra* note 56, at 4 (2022) (allowing inconclusive conclusions in studies "precludes assessment of the performance of forensic examiners for evidence bounded by the two decision criteria, which would be enormously valuable for establishing the true operating characteristics of forensic examiners and error rates of the discipline"); Dror & Scurich, *supra* note 56, at 335, 338 (noting that "error rate studies fall short, and produce inaccurate and misleading error rate estimates" when they do not account for correctness or incorrectness of inconclusive conclusions).

80. *See* Hofmann et al., *supra* note 40, at 323; Monson et al., *supra* note 49, at 3–4; Org. of Sci. Area Comms. for Forensic Sci. Firearms & Toolmarks Subcommittee, *Response to the President's Council of Advisors on Science and Technology (PCAST) Call for Additional References Regarding its Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods"*, at 6 (2016), https://www.nist.gov/organization-scientific-area-committees-forensic-science/firearms-toolmarks-subcommittee [hereinafter *Ensuring Scientific Validity*].

81. *See* Dorfman & Valiant, *supra* note 15, at 3; Hofmann et al., *supra* note 40, at 325; Baldwin et al., *supra* note 49, at 2; Dror & Scurich, *supra* note 56, at 334.

82. PCAST REPORT, *supra* note 2, at 153; Jonothan J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences*, 12 L., PROBABILITY & RISK 89, 95 (2013); Arkes & Koehler, *supra* note 79, at 161.

83. *See, e.g.*, Dorfman & Valiant, *supra* note 15, at 6 ("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 … than the nominal rates coming out of forensic firearms studies so far") (emphasis original); Dror & Scurich, *supra* note 56, at 334 ("errors include reaching an identification (or exclusion) decision where there is insufficient information to justify such a decision; or conversely, reaching an inconclusive decision when there is sufficient information to reach an identification (or exclusion) decision"); Nicholas Scurich, *Inconclusives in firearm error rate studies are not 'a pass'*, 21 L., PROBABILITY & RISK 123, 125 (2022) (opining that until blinded studies with challenging comparisons emerge "firearm error rate studies—much like inconclusive responses—should not be given 'a pass'").

84. *See* A'ssn of Firearm & Toolmark Exam'rs, *supra* note 48.

85. *See* Stanley J. Bajic *et al.*, *Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons*, Technical Report # ISTR-5220, at 240 (2020). It also bears mentioning that, if such inconclusive decisions resulted from the characteristics of fired ammunition as opposed to human bias then computers would likely experience the same issues with eliminations; that turns out not to be the case. *See* Hofmann et al., *supra* note 40, at 344 ("Algorithms generally are symmetric in the assessment of positive and negative criteria—e.g. if a high number of consecutively matching striae is considered evidence in favour of an identification, a low number of consecutively matching striae is consequently evidence in favour of an elimination.").

as inconsequential: taking just one study of bullet comparisons as an example, the false positive rate (depending on the calculation method used) varies from 0.7 percent, to 2 percent, to 66.2 percent.[86]

Abandoning juries to adjudicate such complex debates introduces massive uncertainty into criminal trials, or, as one court put it eloquently decades ago:

> [A] jury of laymen should not, on a case-by-case basis, resolve a dispute in the scientific community . . . [because they are] compelled to make determinations regarding the validity of experimental or novel scientific techniques . . . . [O]ne jury might decide that a particular scientific process is reliable, while another jury might find that the identical process is not . . . . Such inconsistency concerning the admissibility of a given scientific technique or process in criminal cases would be intolerable.[87]

But in vetting firearms examination evidence, courts have split over whether to do just that, and to what extent.[88] On the one hand, as criticism of the field of firearms examination has grown in breadth and nuance, courts not only have divided regarding *Daubert*'s testability, peer review, controlling standards, and general acceptance factors,[89] they have also shown increasing willingness to

86. *See* Monson et al., *supra* note 68, at 89 (20 false positives/2842 total different source conclusions, or 20 false positives/981 false positives and eliminations, or 1881 false positives and inconclusive conclusions/2842 total different source conclusions); *cf.* Amicus Brief, *Abruquah*, *supra* note 77, at 18 (noting as regards the various positions on calculating a false positive rate given inconclusive responses that "even without taking a position as to which most accurately presents the results of the study, it is a significant factor that cannot be ignored").

87. Kansas v. Washington, 622 P.2d 986, 992 (Kan. 1981); *cf.* New York v. Collins, 15 N.Y.S.3d 564, 583 (N.Y. Sup. Ct. 2015) ("[I]f the experts in the DNA field cannot agree on the weight to be given to evidence produced by high sensitivity analysis, it would make no sense to throw such evidence before a lay jury and ask the jurors to give the evidence appropriate weight").

88. For a more thorough history of caselaw regarding the admissibility of firearms examination *see* Garrett et al., *supra* note 1; Brandon L. Garrett, Eric Tucker & Nicholas Scurich, *Judging Firearms Evidence*, 97 S. CAL. L. REV. 101(2024).

89. Multiple courts have concluded that firearms examination fails each of the above *Daubert* factors. *See, e.g.*, United States v. Shipp, 422 F. Supp. 3d 762, 783 (E.D.N.Y. 2019) (considering scholars beyond the bounds of firearms examination practitioners and concluding that "the AFTE Theory has not achieved general acceptance in the relevant community, and this factor weighs against the reliability of the AFTE Theory"); United States v. Adams, 444 F. Supp. 3d 1248, 1264 (D. Or. Mar. 16, 2020) (finding firearms examination "not replicable—and not testable—because it cannot be explained in a way that would allow an uninitiated person to perform the same test in the same way"); United States v. Tibbs, No. 2016-CF1-19431, 2019 WL 4359486, at *10, 20 (D.C. Super. Sep. 5, 2019) (finding controlling standards factor weighed against admissibility and saying of the AFTE *Theory of Identification* "under this so-called standard, the process for determining what constitutes a 'match' lacks defined criteria; it is merely unconstrained subjectivity masquerading as objectivity" and similarly concluding as to peer review that said factor "on its own does not, despite the sheer number of studies conducted and published, work strongly in favor of admission of firearms and toolmark identification testimony"). Other courts, however, have found that *Daubert*'s factors weigh in favor of admissibility. *See, e.g.*, United States v. Felix, 77 V.I. 714, 2022 WL 17250458, at *13 (D.V.I. Nov. 28, 2022) (expressing concern about peer review within the field of firearms examination, but in considering outside scrutiny of the field, finding "that the AFTE methodology has been subjected to sufficient peer review and publication"); United States v. Rhodes, No. 3:19-CR-00333-MC, 2023 WL 196174, at *6 (D. Or. Jan. 17, 2023) ("[T]he weight of authority suggests that the AFTE method does enjoy general acceptance in the relevant scientific community—forensic ballistic examiners."); United States v. Romero-Lobato, 379 F. Supp. 3d 1111, 1118 (D. Nev. 2019)

circumscribe the testimony of practitioners.[90] On the other hand, however, most of the limitations imposed by courts are unlikely to impact juror assessment of firearms comparison testimony,[91] and judicial faith in false positive rates for the field within the low single digits has remained curiously resolute.[92] That is not to say that courts have been wholly insensitive to the criticisms of outside scholars; to the contrary, multiple judges have refused to rely on set-based studies, expressed concerns regarding inconclusive responses, and rejected the reliability of error rate estimates outright.[93]

("[T]he testability element is a key question in determining whether expert testimony should be admitted. There is little doubt that the AFTE method of identifying firearms satisfies this *Daubert* element.") (internal quotations and citations omitted); *Otero*, 849 F. Supp. 2d at 435 ("the maintenance of industry-compliant standards by the NJSP for conducting a firearms and toolmark identification examination…further support the reliability and therefore admissibility of the expert testimony").

90. *See, e.g.*, Mark Page, Jane Taylor & Matt Blenkin, *Forensic Identification Science Evidence Since* Daubert*: Part I-A Quantitative Analysis of the Exclusion of Forensic Science Evidence*, 56 J. FORENSIC SCIS. 1180, 1182 (2011) (identifying total thirty-seven challenges firearms examination testimony that resulted in either exclusion or limitation of the proffered evidence with reliability as the reason for exclusion in 20 of those); United States v. Mouzone, 696 F.Supp.2d 536, 569, 572–73 (D. Maryland 2009) (concluding that neither conclusions of absolute nor practical certainty of a match were factually warranted); United States v. Taylor, 663 F.Supp.2d 1170, 1180 (D. NM 2009) ("[B]ecause of the limitations on the reliability of firearms identification evidence discussed above, Mr. Nichols will not be permitted to testify that his methodology allows him to reach this conclusion as a matter of scientific certainty. Mr. Nichols also will not be allowed to testify that he can conclude that there is a match to the exclusion, either practical or absolute, of all other guns."); United States v. Green, 405 F. Supp. 2d 104, 124 (D. Mass. 2005) (permitting testimony regarding observations but no ultimate opinion about source); United States v. Glynn, 578 F. Supp. 2d 567 (S.D.N.Y. 2008) (noting that, given the lack of data supporting the discipline "ballistics lacked the rigor of science," and limiting testimony of match to a conclusion of "more likely than not" instead of even "reasonable ballistics certainty" to ensure that "a conviction in a criminal case may not rest *exclusively* on ballistics testimony"); United States v. Monteiro, 407 F. Supp. 2d 351, 375 (D. Mass. 2006) (limiting testimony to "reasonable degree of ballistic certainty"); *Diaz*, 2007 U.S. Dist. LEXIS 13152, at *41–42 (precluding matches to the exclusion of all other guns in the world); *Tibbs*, No. 2016-CF1-19431, 2019 WL 4359486, at *21–22 (restricting testimony to cannot exclude language); *Shipp*, 422 F. Supp. 3d at 766 limiting to "consistent with"); *Felix*, 77 V.I. 714, 2022 WL 17250458, at *17 (same).

91. *See* Brandon L. Garrett, Nicholas Scurich & William E. Crozier, *Mock Jurors' Evaluation of Firearm Examiner Testimony*, 44 LAW & HUM. BEHAV. 412, 413 (2020) (conducting experiment vetting lay reactions to variations of firearms examiner testimony and concluding that only limitations of "cannot exclude" had any significant impact).

92. *See, e.g.*, United States v. Harris, 502 F. Supp. 3d 28, 39 (D.D.C. 2020); *Romero-Lobato*, 379 F. Supp. 3d at 119–20; United States v. Johnson, No. (S5) 16 CR. 281, 2019 WL 1130258, at *19 (S.D.N.Y. Mar. 11, 2019); United States v. Hunt, 464 F. Supp. 3d 1252, 1258 (W.D. Ok. 2020); *Felix*, 77 V.I. 714, 2022 WL 17250458, at *17; United States v. Chavez, No. 15-CR-00285-LHK-1, 2021 WL 5882466, at *4 (N.D. Cal. Dec. 13, 2021); United States v. Pete, No. 3:22CR48-TKW, 2023 WL 4928523, at *4 (N.D. Fla. July 21, 2023); United States v. Blackman, No. 18-CR-00728, 2023 WL 3440384, at *6 (N.D. Ill. May 12, 2023); United States v. Brown, 973 F.3d 667, 704 (7th Cir. 2020); *Rhodes*, No. 3:19-CR-00333-MC, 2023 WL 196174, at *4; United States v. Gist-Holden, 629 F. Supp. 3d 841, 846 (N.D. Ind. 2022); United States v. Dunham, 654 F. Supp. 3d 1183, 1191 (E.D. Okla. 2023).

93. *See* United States v. Briscoe, No. 20-CR-1777 MV, 2023 WL 8096886, at *9 (D.N.M. Nov. 21, 2023) ("Even the error rates reported in black-box studies of toolmark analysis are questionable, as many studies count inconclusive responses as correct without explanation or justification"); United States v. Cloud, 576 F. Supp. 3d 827, 843 (E.D. Wash. 2021) ("But providing examiners in the study setting the option to essentially "pass" on a question, when the reality is that there is a correct answer—the casing either was or was not fired from the reference firearm— fundamentally undermines the study's analysis of the methodology's foundational validity and that of the error rate."); *Felix*, No. CR 2020-0002, 2022 WL 17250458, at *16 (refusing to rely on set-based

But through these divides, when the time has come to cite a false positive rate, courts from across the aisle in terms of ultimate admissibility decisions have nevertheless (and in the face of all the extensive scientific criticism outlined herein) been nearly uniform: The judges in *Shipp* and *Adams* may have considered firearms examination false positive rates too high to favor admissibility,[94] the judge in *Felix* "significant but not so high" as to weigh against the field,[95] and others "low [and] acceptable under *Daubert*," but all would place the relevant figure as hovering at, or below, 2 percent.[96] Indeed, despite thorough skepticism of existing validation studies of firearms examination, even the Maryland Supreme Court in *Abruquah* conceded that "[t]he relatively low rate of 'false positive' responses in studies conducted to date is the most persuasive piece of evidence in favor of admissibility of firearms identification evidence."[97] If existing critiques have not dislodged such uncritical acceptance of misidentification rates, then a new approach must be added to the pile, because all these courts—the doubting and the credulous alike—have been deceived. False positive rates in firearms examination studies have repeatedly exceeded 2 percent, and where they have not, the accuracy observed likely turns far more on the simplicity of study comparisons than the foundational validity of firearms examination methods. It is to those claims that this Article now turns, first by elucidating the importance of testing challenging comparisons, and second by illuminating the field of firearms examination's failure to do so without provoking truly disturbing rates of misidentification.

---

studies or to use calculations counting inconclusive responses as correct); *Tibbs*, No. 2016-CF1-19431, 2019 WL 4359486, at *13–18 (explaining why use of set-based studies, test-taking bias, and the issue of inconclusive conclusions all undermine the reliability of false positive estimates"); Illinois v. Winfield, No. 15CR14066-01, at 23–24 (Cir. Ct. Cook Cnty. Feb. 8, 2023) (on file with author) (discussing the lack of reliability in false positive rates due to inconclusive responses as well as the negating impact on validity of repeatability and reproducibility figures for the field).

94. *See Shipp*, 422 F. Supp. 3d at 778 ("the[s]tudy that most closely resembles fieldwork estimated that a firearms toolmark examiner may incorrectly conclude that a recovered piece of ballistics evidence matches a test fire once out of every 46 examinations. When compared to the error rates of other branches of forensic science—as rare as 1 in 10 billion for single source or simple mixture DNA comparisons—this error rate cautions against the reliability of the AFTE Theory"); *Adams*, 444 F. Supp. 3d at 1265 ("It is possible that the error rate for toolmark testing is very low, but it is more likely that it is not. Assuming false positive test results lead to wrongful convictions, a wrongful conviction rate of 1 in 46 is far too high. The best test results would favor the government, but it is unlikely those tests reflect real-world error rates. The worst results favor Defendant. At most, then, this factor of the *Daubert* test is neutral as to both parties. In my opinion, it cuts somewhat in favor of Defendant."); Oregon v. Moore, No. 18CR77176, at 24 (Cir. Ct. Or. Aug. 8, 2023) ("[T]here was like a really low rate of false positives . . . . But still, you know, the false positive rate doesn't need to be very high before it's really problematic when it comes to scientific evidence. And especially, you know, you're in the criminal justice system and you have a false positive rate of two percent or something, that can be really problematic.").

95. *Felix*, No. CR 2020-0002, 2022 WL 17250458, at *17.

96. *Pete*, No. 3:22CR48-TKW, 2023 WL 4928523, at *4; *see also Chavez*, No. 15-CR-00285-LHK-1, 2021 WL 5882466, at *4 ("The Court finds this factor weighs slightly in favor of admissibility for two reasons. First, the weight of authority suggests the potential error rate is between 0–1%. Second, even if the error rate is as large as 2.2%, the Court disagrees with the conclusion by the *Adams* and *Shipp* courts that such an error rate is impermissibly high.").

97. 296 A.3d 961, 687 (going on to say that "[o]n balance, however, the record does not demonstrate that that rate is reliable, especially when it comes to actual casework").

III. TESTING THE SPECTRUM, COVERING THE FACTOR SPACE, AND INCLUDING CHALLENGING COMPARISONS

Anyone serving as a judge in the United States almost certainly underwent a slew of testing during their educational journey towards a law degree including, one hopes, exams with truly vexing questions designed to test the limits of understanding. Their literature finals undoubtedly asked more than which character in *Hamlet* utters the words "to be, or not to be;" their performance when taking math classes was almost definitely vetted beyond an ability to divide nine by three; and passing the bar surely required knowledge exceeding whether police (generally) must obtain a warrant to search a private home. But despite these experiences, courts have nigh-exclusively declined to ask, or been untroubled by, whether accuracy studies of firearms examination include samples sufficiently varied in their complexity and difficulty to estimate the potential for error across the range of circumstances expected in casework.[98] Nevertheless—and though too often forgotten by both sides of the debate on the admissibility of firearms examination evidence given a focus on the *number* of studies conducted[99]—appropriate method validation necessitates that researchers "push the system until it fails in order to understand the potential limitations."[100] In other words, if courts are to reach defensible conclusions about the validity of firearms examination methods, they must be based on studies with samples "represent[ing] the full range and distribution of types and difficulty normally seen in casework."[101] And if we are to grant practitioners the

98. *See, e.g.*, United States v. Romero-Lobato, 379 F. Supp. 3d 1111, 1119–20 (D. Nev. 2019) (declining to adopt "strict requirement[s] for which studies are proper and which are not" and finding that "low" rates of error favor the admissibility of firearms examination); United States v. Monteiro, 407 F. Supp. 2d 351, 367–68 (D. Mass. 2006) (expressing some concern that "results might instead indicate that the test was somewhat elementary" but immediately transitioning to explain that "there is no evidence that the tests are inaccurate or otherwise deficient" and thus "the government has established that known error rate is not unacceptably high").

99. *Compare* Eric S. Lander, *Fixing Rule 702: The PCAST Report and Steps to Ensure the Reliability of Feature-Comparison Methods in the Criminal Courts*, 86 FORDHAM L. REV. 1661, 1672 (2018) ("With only a single well-designed study estimating accuracy, PCAST judged that firearms analysis fell just short of the criteria for scientific validity, which requires reproducibility. A second study would solve this problem"); *with* Valerio & Bunn, *supra* note 70 ("That second study has been done, as well as several others that meet PCAST's prescribed standards and vindicate firearms identification. The time has arrived for the scientific and legal communities to recognize its reliability in shooting investigations"); *but see* Scurich et al., *supra* note 26, at 7 (calling it "scientifically naive to say that once a second study was completed, the work of the field was done").

100. John Butler, *NIST DNA Analysis Webinar Series: Validation Concepts and Resources- Part I Validation Overview*, NAT'L INST. OF STANDARDS AND TECH. (Aug. 6, 2014), www.nist.gov/system/files/documents/forensics/01_ValidationWebinar-Butler-Aug2014.pdf; *see also* NAT'L INST. OF FORENSIC SCI. OF THE AUSTL. N.Z. POLICING ADVISORY AGENCY, EMPIRICAL STUDY DESIGN IN FORENSIC SCIENCE: A GUIDELINE TO FORENSIC FUNDAMENTALS 10 (2019) ("A validation is not a test for 100% performance; it is a tool to determine when a method works and when it does not. In fact, evidence of 100% correct responses is an indication that the test materials were not sufficiently complex. The experimental design should have considered the range of outcomes possible to ensure that the outer bounds of the claim are assessed.") [hereinafter GUIDELINE TO FORENSIC FUNDAMENTALS].

101. OSAC HUMAN FACTORS REPORT, *supra* note 16, at 11; *see also* PCAST REPORT, *supra* note 2, at 52 ("[S]tudies must involve a sufficiently large number of examiners and must be based on sufficiently *large*

powers and privileges unique in our legal system to "expert" witnesses,[102] then empirical evidence should assure us that firearms examiners indeed display a crucial "hallmark of expertise," namely, "[t]he ability to differentiate between similar but not identical stimuli" (i.e. close non-matches).[103]

Such insights—far from novel, ambitious, or controversial—pervade scientific literature and standards within and outside of forensics. In fact, researchers in the realm of diagnostic testing realized in the 1970s that failure to account for performance across populations (including challenging cases) had led to an unfortunate cycle of "early optimism" and "subsequent disillusionment" with screening tests as critical as those for cancer.[104] Since then, this problem of "spectrum bias" (sometimes also called "spectrum effect") has received substantial attention, with industry testing standards warning that "[e]liminating . . . difficult cases produces an overly optimistic picture" of method performance,[105] meta-studies finding hundreds of examples of

collections of *known* and *representative* samples from relevant populations to reflect the range of features or combinations of features that will occur in the application.") (emphasis original); GUIDELINE TO FORENSIC FUNDAMENTALS, *supra* note 100, at 9–10 ("It is also important to ensure that the test materials used reflect the range of materials and difficulty encountered in casework and that conditions are consistent with those in an operational setting . . . testing only complete, high quality samples will not explore the accuracy of the method on partial, distorted and degraded material."); Garrett et al., *supra* note 88, at 144–45 ("[T]here are also questions about whether the materials being used in the studies, such as the types of firearms and the quality of the fired items, are sufficiently representative to draw inferences about the field writ large. By design, studies should be of varying degrees of difficulty . . . ."); Itiel E. Dror, *The Error in "Error Rate": Why Error Rates Are So Needed, Yet So Elusive*, 65 J. FORENSIC SCIS. 1034, 1035 (2020) ("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.").

102. While "expert" witnesses "may testify in the form of an opinion or otherwise," FED. R. EVID. 702, lay witnesses can do so only in limited circumstances, FED. R. EVID. 701.

103. David J. Weiss & James Shanteau, *Empirical Assessment of Expertise*, 45 HUM. FACTORS 104, 107 (2003). In fact, when researchers in other pattern-matching fields have set out to establish the existence of expertise (firearms examination, though the topic falls somewhat beyond the aims of this article, has never demonstrated that its purported "experts" actually possess skills beyond those of lay people in the comparison of fired bullets and cartridge cases) they have found that performance only truly differs between experts and novices when it comes to distinguishing between non-mated bears of prints that bear substantial coincidental similarity. *See, e.g.*, Matthew B. Thompson & Jason M. Tangen, *Human Matching Performance of Genuine Crime Scene Latent Fingerprints*, 38 LAW & HUM. BEHAV. 84, 87, 91 (2014) ("The superior performance of experts in this experiment was not simply a function of their ability to match prints, per se, but a result of their ability to identify highly similar, but nonmatching fingerprints as such.").

104. *See* D. F. Ransohoff & A. R. Feinstein, *Problems of Spectrum and Bias in Evaluating the Efficacy of Diagnostic Tests*, 17 N. ENGL. J. MED. 926 (1978).

105. U.S. FOOD & DRUG ADMIN., STATISTICAL GUIDANCE ON REPORTING RESULTS FROM STUDIES EVALUATING DIAGNOSTIC TESTS 13 (2007) ("Estimates of diagnostic accuracy are subject to spectrum bias when the subjects included in the study do not include the complete spectrum of patient characteristics; that is, important patient subgroups are missing. For example, there are studies that include only very healthy subjects and subjects with severe disease, omitting the intermediate and typically more difficult cases to diagnose. The accuracy measures reported from these studies are subject to spectrum bias.") (internal citations omitted); CLINICAL & LAB'Y STANDARDS INST., EP12-ED3: EVALUATION OF QUALITATIVE, BINARY OUTPUT EXAMINATION PERFORMANCE 43 (2023) ("Samples included in a clinical performance study should represent the intended-use population. If the subjects enrolled are not representative of the intended-use population, estimates of clinical performance are subject to a spectrum effect. For example, if only subjects from the extreme

"diagnostic studies with methodological shortcomings . . . [that] overestimate the accuracy of a diagnostic test, particularly those including nonrepresentative patients,"[106] and clinicians cautioning physicians against overreliance on limited testing data.[107] So, too, has the forensic sphere (even beyond the general exhortations noted above)[108] embraced the need for attention to testing representative samples (including difficult cases). Taking DNA as an example, not only does a leading textbook caution that "laboratories cannot adequately understand performance characteristics of low-template, complex DNA mixtures from having run a few high-template, simple DNA mixtures,"[109] but when the National Institute of Standards and Technology (NIST) set out to vet the foundational validity of contemporary mixture interpretation approaches, it demanded that empirical data show acceptable performance across the "factor space," defined as "the totality of scenarios and associated variables (factors) that are considered likely to occur in actual casework."[110] Indeed, despite the existence of multiple studies testing thousands of comparisons involving different contributor numbers and template amounts,[111] NIST would ultimately conclude, using its factor space approach, that "there is not enough publicly

ends of the TC are sampled (eg, either healthy subjects or subjects with advanced-stage disease), performance can appear to be better than it is. This effect happens because subjects with results near the underlying examination cutoff that are omitted tend to be more difficult to diagnose correctly. When subjects are closer to the underlying cutoff, the inherent variability of the examination can increase the chance of an incorrect diagnosis if the examination has not been correctly designed.").

106. *See* Jeroen G. Lijmer, B. W. Mol, S. Heisterkamp, G. J. Bonsel, M. H. Prins, J. H. van der Meulen & P. M. Bossuyt, *Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests*, 282 JAMA 1061 (1999) (detailing a meta study of 184 original studies evaluating 218 diagnostic tests, which found that "[t]hese data provide empirical evidence that diagnostic studies with methodological shortcomings may overestimate the accuracy of a diagnostic test, particularly those including nonrepresentative patients"); Anne W. S. Rutjes, Johannes B. Reitsma, Marcello Di Nisio, Nynke Smidt, Jeroen C. van Rijn & Patrick M. M. Bossuyt, *Evidence of Bias and Variation in Diagnostic Accuracy Studies*, 174 CMAJ 469 (2006) (presenting meta-analyses with 487 primary studies of test evaluations found major overestimation of accuracy based on exclusion of "complex cases" and the inclusion of healthy controls which are simple to diagnose and thereby lower false positives).

107. *See* Brian H. Willis, *Spectrum Bias—Why Clinicians Need To Be Cautious When Applying Diagnostic Test Studies*, 25 FAM. PRAC. 390 (2008).

108. *See supra* notes 16, 100-101.

109. JOHN M. BUTLER, ADVANCED TOPICS IN FORENSIC DNA TYPING: INTERPRETATION 164–66 (1st ed. 2010).

110. John M. Butler, Hari Iyer, Rich Press, Melissa K. Taylor, Peter M. Vallone & Sheila Willis, *NISTIR 8351-DRAFT DNA Mixture Interpretation: A NIST Scientific Foundation Review*, NAT'L INST. STANDARDS & TECH., 60–61 (2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8351-draft.pdf.

111. *See, e.g.*, Tamyra R. Moretti, Rebecca S. Just, Susannah C. Kehl, Leah E. Willis, John S. Buckleton, Jo-Anne Bright, Duncan A. Taylor & Anthony J. Onorato, *Internal validation of STRmix™ for the Interpretation of Single Source and Mixed DNA Profiles*, 29 FORENSIC SCI. INT'L: GENETICS 126 (2017); Jo-Anne Bright et al., *Internal Validation of STRmix™ – A Multi Laboratory Response to PCAST*, 34 FORENSIC SCI. INT'L: GENETICS 11 (2018).

available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation practices."[112]

Establishing the appropriate range of samples to test for fields (like firearms examination) that lack objective metrics of difficulty is not trivial[113] and requires elaborate and nuanced research.[114] The PCAST (President's Council of Advisors on Science and Technology) Report, case in point, articulated detailed guidance on the scope of data available and the limits of validity for DNA (looking at contribution number, the ratio between donors, and the overall quantity of genetic material) but seemingly could not for disciplines like latent print comparison and firearms examination.[115] But scholars have nevertheless keyed in on one essential component of appropriately representative testing for different source comparisons: the inclusion of close non-matches (i.e. samples originating from different sources that bear significant coincidental similarities and few distinguishing differences).[116] Without statistical foundations for estimations of the rarity of particulars arrangements of features, the frequency at which such close non-matches will confront examiners cannot be known,[117] but their potential to provoke false positives should not be understated: the coincidental similarity between a print associated with a train bombing in Madrid and those of Brandon Mayfield precipitated perhaps the most infamous

112. Butler et al., *supra* note 110, at 75.

113. *See* Dror, *supra* note 101, at 1035 ("[O]ne 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.").

114. *See* OSAC HUMAN FACTORS REPORT, *supra* note 16, at 12 (emphasizing the importance of research "to assess in a rigorous manner the difficulty of the analytic results examiners must routinely reach"); Phillip J. Kellman, Jennifer L. Mnookin, Gennady Erlikhman, Patrick Garrigan, Tandra Ghose, Everett Mettler, David Charlton & Itiel E. Dror, *Forensic Comparison and Matching of Fingerprints: Using Quantitative Image Measures for Estimating Error Rates Through Understanding and Predicting Difficulty*, 9 PLOS ONE e94617 (2014) (doing so in the context of latent print examination).

115. *Compare* PCAST REPORT, *supra* note 2, at 75–83 *with id.* at 87–114.

116. *See* Johnathan J. Koehler & Shiquan Liu, *Fingerprint Error Rate on Close Non-Matches*, 66 J. FORENSIC SCIS. 129, 130 (2021) (defining close non-matches as items which "have many common features and few discernible dissimilar features'); *infra* notes 120-122; Heidi Eldridge, Marco De Donno, Margaux Girod & Christophe Champod, *Coping with Close Non-Matches in Latent Print Comparison (re-)Training*, at 2 (2022) https://www.ojp.gov/pdffiles1/nij/grants/305757.pdf ("CNM prints are of crucial importance in developing the expertise of latent print examiners . . . because they constitute the worst-case scenario for a comparison between impressions originating from different sources."); Scurich, *supra* note 83, at 125 (2022) ("researchers should intentionally select challenging test items"); Dror, *supra* note 101, at 1035 ("[N]onmatches need not only be ground truth nonmatches, but also need to include challenging and difficult cases, 'look alikes' that are nevertheless a nonmatch."); Thompson & Tangen, *supra* note 102, at 88 (emphasizing the need to test close non-matches because "[d]istinguishing such highly similar, but nonmatching, print pairs from actual matching print pairs is potentially the most difficult task that fingerprint examiners face"). *See also* Trans. of Proceedings, Illinois v. Winfield, 15CR14066-01, at 205, 211 (Cir. Ct. Cook Cnty. Mar. 1, 2022). Todd Weller, a firearms examiner conceded as much in a recent hearing: "Q. When you are testing specificity, it is important to also make sure that you are including different source comparisons that bear some level of coincidental similarity, right? A. Yes. Q. It is important, in other words, to try and test close nonmatches? A. Yes." *Id.*

117. *See supra* note 35; NAS REPORT, *supra* note 2, at 153–54; BALLISTIC IMAGING, *supra* note 33, at 3–4; Schwartz, *supra* note 31, at 12–13, 20–21.

misidentification in the world of pattern-matching forensics,[118] and studies focused in on the most difficult close non-matches in the realms of DNA and latent print comparison have generated truly harrowing rates of error.[119] In keeping with such concerns, accuracy studies from the latent print realm not only have emphasized that "[w]ithout deliberately sought-out close non-match distractors, it is highly unlikely that … different source trials presen[t] a meaningful challenge,"[120] they have also gone to great lengths to seek out and include such comparisons (trolling available databases for coincidentally similar pairs, and even assigning subject matter experts to select database pairs with substantial coincidental similarity).[121] But while firearms examiners, and other researchers exploring the field's accuracy, have often paid lip service to such ideals—either by acknowledging that "[i]t is important to test the limits of examiners using the most challenging conditions for the method evaluated, in order to get a true picture of error rates,"[122] or by claiming to have chosen firearms "for their propensity to produce challenging and ambiguous test specimens, creating difficult comparisons for examiners"[123]—the critical question whether existing studies have addressed such samples in their estimation of error rates (whether they have included close non-matches that

118. *See* U.S. DEPT. OF JUST., OFF. OF THE INSPECTOR GENERAL, A REVIEW OF THE FBI'S HANDLING OF THE BRANDON MAYFIELD CASE 7 (2006).

119. *See* John M. Butler, *NIST Interlaboratory Studies Involving DNA Mixtures (MIX05 and MIX13): Variation Observed and Lessons Learned*, 37 FORENSIC SCI. INT'L: GENETICS 81, 87(2018) (describing how 74 of 108 laboratories, or 68.5 percent falsely included one person of interest as a contributor to the study's most difficult DNA mixture); Koehler & Liu, *supra* note 116, at 129 (reporting false positive rates of 15.9 percent and 28.1 percent on two especially challenging latent print close non-matches, and emphasizing that "[c]oncern about false identifications from database-derived CNMs is not merely theoretical. The Brandon Mayfield case . . . shows both that database CNMs exist and may fool even the best fingerprint examiners.").

120. Heidi Eldridge, Marco De Donno & Christophe Champod, *Testing the Accuracy and Reliability of Palmar Friction Ridge Comparisons – A Black Box Study*, 318 FORENSIC SCI. INT'L 110457, at 1 (2021).

121. *See id.* at 2; Bradford T. Ulery, R. Austin Hicklin, JoAnn Buscaglia & Maria Antonia Roberts, *Accuracy and Reliability of Forensic Latent Fingerprint Decisions*, 108 PROC. NAT'L ACADEMY SCIS. 7733, 7734 & SI Appendix 4–5 (2011); Thompson & Tangen, *supra* note 102, at 86; *see also* Dror, *supra* note 113, at 1035 (recommending the use of databases for the fields of fingerprint and firearms comparison to seek out challenging comparison sets for validation). That is not to say that false positive rates generated in studies of latent print examination are beyond scrutiny. *See generally* Ralph N. Haber & Lyn Haber, *Experimental Results of Fingerprint Comparison Validity and Reliability: A Review and Critical Analysis*, 54 SCI. & JUST. 375 (2014). Indeed, much works remains to reconcile the low error rates of some studies and the double-digit false positive rates generated in others. *See* Koehler & Liu, *supra* note 116, at 133. But at least the field of latent print comparison has begun that work in earnest.

122. Best & Gardner, *supra* note 15, at 28; *see* Eric F. Law & Keith B. Morris, *Evaluating Firearm Examiner Conclusion Variability Using Cartridge Case Reproductions*, 66 J. FORENSIC SCIS. 1704, 1705 (2021) ("[I]t is important to select a group of firearms that are representative of what firearm examiners encounter in casework."); Chad Chapnick, Todd J. Weller, Pierre Duez, Eric Meschke, John Marshall & Ryan Lilien, *Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for Firearm Forensics*, 66 J. FORENSIC SCIS. 557, 559 (2021) ("The power of error rate studies and validation studies is related to the breadth and complexity of specimens included.").

123. Monson et al., *supra* note 68, at 87.

might actually stand a chance of establishing the limits of their method) remained largely unexplored in scholarly literature, at least until now.[124]

IV. THE KIDDY STUFF PERMEATING FIREARMS EXAMINATION VALIDATION STUDIES AND THE EXCEPTIONS THAT BREAK THE RULE

Concluding that the field of firearms examination has not explored the full range of circumstances and difficulty possible in casework scarcely requires more than some simple arithmetic. Just counting the number and variety of samples tested nearly suffices on its own given that researchers exploring the accuracy of the field have utilized an embarrassingly paltry sample of the hundreds of millions of guns circulating in the United States,[125] not to mention of the great diversity (in terms of calibers, manufacturers, and makes and models) of firearms recovered in relation to gun crimes.[126] Indeed, the four pairwise studies which have generated the types of false positive rates cited by most courts have used just one caliber (9mm Luger), two manufacturers, and thirty-eight total barrels for their bullet comparisons,[127] and just two calibers (9mm Luger and 40 Smith & Weston), six manufacturers, and ninety-eight guns for their cartridge case comparisons.[128] Thus, before even accounting for the challenge-level of the samples included in accuracy studies of firearms examination to date, the negligible range and diversity of gun variables tested cautions against regarding the record underlying the field sufficient to establish its validity.[129]

Things, however, come into ever sharper resolution when considering the former and asking whether existing studies have included samples bearing

124. *See* Garrett et al., *supra* note 88, at 145 ("Unlike other forensic identification fields, none of these studies [of firearms examination] have used technology or databases to ensure the test items are challenging. Nor has there been any careful analysis of how representative or challenging these studies are, and this basic problem has not received the judicial attention that it should.").

125. *See* Christopher Ingraham, *There Are More Guns Than People in the United States, According to a New Study of Global Firearm Ownership*, WASH. POST (June 19, 2018), https://www.washingtonpost.com/news/wonk/wp/2018/06/19/there-are-more-guns-than-people-in-the-united-states-according-to-a-new-study-of-global-firearm-ownership (reporting 393 million guns in the United States).

126. *See* BUREAU OF ALCOHOL, TOBACCO, & FIREARMS, NATIONAL FIREARMS COMMERCE & TRAFFICKING ASSESSMENT, VOLUME II: CRIME GUN INTELLIGENCE & ANALYSIS, PART III: CRIME GUNS RECOVERED & TRACED WITHIN THE UNITED STATES AND ITS TERRITORIES 18–22 (2023), https://www.atf.gov/firearms/national-firearms-commerce-and-trafficking-assessment-nfcta-crime-guns-volume-two [hereinafter BATF: VOLUME II, PART III]; Sarah Kollmorgan, *Chicago Criminals' Favorite Gunmakers: A Visual Ranking*, TRACE (Jan. 6, 2016), https://www.thetrace.org/2016/01/chicago-crime-guns-chart; CAL. DEP'T OF JUST. DIV. OF L. ENFORCEMENT BUREAU OF FORENSIC SERVICES, FIREARMS USED IN THE COMMISSION OF CRIMES (2020), https://oag.ca.gov/sites/default/files/firearms-report-20.pdf; U.S. DEP'T JUST. BUREAU OF STAT., GUNS USED IN CRIME (1995), https://bjs.ojp.gov/content/pub/pdf/GUIC.pdf.

127. *See* Monson et al., *supra* note 68, at 88.

128. *See id.*; Guyll et al., *supra* note 68, at 9; Baldwin et al., *supra* note 68, at 3; M. A. Keisler, S. Hartman, A. Kilmon & M. Oberg, *Isolated Pairs Research Study*, 50 AFTE J. 56, 56–57 (2018).

129. *See generally* Spiegelman & Tobin, *supra* note 59 (criticizing earlier studies of firearms examination for their limited sampling of gun types and manufacturing variables); Dorfman & Valiant, *supra* note 15, at 5 ("With a few exceptions, each of the forensic firearms studies to date focuses on a single firearm. This gives rise

coincidental similarity likely to, or capable of, provoking misidentifications and adequately estimating false positive rates. As discussed more fully below,[130] comparing fired bullets and cartridges comes with more than its fair share if difficulties. Yet, hand-in-hand with its focus, not on accuracy more generally, but instead on providing law enforcement with as many matches as possible—the field's governing standard (to the extent it could be called one) is, after all, titled the *Theory of Identification*, not of source conclusions, comparison criteria, or some other more neutral phrasing—firearms examination has largely refused to explore performance on such challenging samples. So deeply has a bias in favor of testing matching accuracy to the exclusion of elimination accuracy leached into existing studies that one group of researchers felt comfortable with the frankly audacious level of honesty required to admit that their study involved "no intention to select the pairs in the true elimination sets that would attempt to lead the participants into making a false positive source-attribution conclusion (*e.g.*, strong carry-over of subclass characteristics between the pairs)."[131] Unfortunately, while such transparency may buck convention, similar failures to vary difficulty, and thereby cover the range of challenges presented by casework, have been the norm.

That much should be obvious from even a cursory review of the characteristics of, and false positive rates generated by, the seven sample-to-sample studies (those appropriately designed to unambiguously allow for false positive rate calculations)[132] thus far conducted on the accuracy of bullet and cartridge case comparisons and summarized in Table 1.[133] That table provides

---

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."). Indeed, scholars have aptly demonstrated the problem of expecting minimal sampling to have uncovered, and thereby accounted for in existing testing, the most difficult cases examiners may well still encounter in casework with an apt statistical hypothetical:

> Suppose that exactly 100 pairs of firearms out of an estimated 100,000 guns in a Texas town share indistinguishable gun barrel markings. If each of 100 firearms experts examined 10 pairs of guns from the town's gun population every day for 10 years…there is about a 93% chance that none of the indistinguishable pairs will have come under examination. That is, despite 1,000 'collective years' of forensic science experience…the failure to find even a single pair of guns with indistinguishable markings would offer little basis for drawing conclusions about whether gun barrel markings, even in this single town, are unique.

Michael J. Saks & Jonothan J. Koehler, *The Individualization Fallacy in Forensic Science Evidence*, 6 VANDERBILT L. REV. 199, 213 (2008).

130. *See infra* Part IV.A.

131. Knowles et al., *supra* note 49, at 517.

132. This summary, and the focus on pairwise studies more generally, has the effect of excluding one of the studies showing substantial rates of error mentioned in the introduction because it utilized an open-set design. *See* Knapp & Garvin, *supra* note 15.

133. The discussion of firearms examination validation studies in this Part (and Table 1), technically exclude three sample-to-sample studies because various choices made by their designers preclude direct comparison to the remaining record underlying the field. *See* Law & Morris, *supra* note 122; Chapnick et al., *supra* note 122; Erwin J. A. T. Mattijssen, Cilia L. M. Witteman, Charles E. H. Berger, Xiaoyu A. Zheng, Johannes A. Soons &

false positives calculated by the two approaches that *proponents* of firearms examiners have deemed informative[134] in an effort, again, to meet judicial opinions where they lay and avoid unnecessarily tethering the arguments of this article with other conversations about the trustworthiness of error rates for the field. It includes confidence intervals to establish a plausible "range of values that are reasonably compatible with the results" of each study given that the limited samples of any individual research effort "cannot provide 'exact' values," they can merely estimate accuracy rates.[135] And it focuses only on within-class comparisons because the alternative (encompassing class eliminations) "would not be meaningful, as it is assumed that firearm examiners would be able to separate cartridge cases fired from firearms with different class characteristics."[136] Of the studies summarized in Table 1, none utilize databases to seek out close non-matches, three report no efforts whatsoever to done so,[137] and another two merely assumed they could cover the full range of difficulty possible in casework by relying on a small number of consecutively manufactured guns.[138] Worse, when researchers (all European, to the shame of the firearms examination community here in the United States) have employed

Reinoud D. Stoel, *Firearm Examination: Examiner Judgments and Computer-Based Comparisons*, 66 J. FORENSIC SCIS. 96 (2021). The first excluded a participant (and thus did not report the full range of that individual's conclusions) who had completed the entire sample-to-sample portion of the study. *See* Law & Morris, *supra* note 122, at 1709. The second involved the use of a technology that likely enhances examiner abilities beyond traditional methods by providing access to "visual detail beyond what is typically seen with traditional Light Comparison Microscopy." Cadre Forensics, *Cadre-VCM Validated Virtual Comparison Microscopy (VCM)*, (last accessed Dec. 20, 2023), https://www.cadreforensics.com/VirtualComparisonMicroscopy.html. And the third artificially separated out performance on breech face and firing pin comparisons as opposed to overall cartridge case comparison accuracy. *See* Mattijssen et al., *supra*, at 103. Excluding these studies does not change the analysis in this article and, in fact, is conservative relative to the point that false positive rates for firearms examination have exceeded the 2 percent figure commonly cited by judges. One of those studies (which generated a false positive rate in line with judicial views) clearly did not include challenging comparisons. *See* Chapnick et al., *supra* note 122; *infra* note 206. And the other two resulted in false positive rates multiple times larger than those cited by the courts: the upper bound for one (when factoring in the excluded participant and assuming the best possible performance on that individual's part) is 7.6 percent on conclusive decisions, *see* Law & Morris, *supra* note 122, at 1709–10 (6 false identifications, 155 known eliminations, 8 eliminations assumed for excluded participant), and the false positive rates for breech face and firing pin comparisons in the second were 11.2 percent and 12.1 percent, respectively (again looking at conclusive decisions), *see* Mattijssen et al., *supra*, at 103.

134. *Ensuring Scientific Validity*, *supra* note 80, at 6.

135. PCAST REPORT, *supra* note 2, at 51, 152–53; NAS REPORT, *supra* note 2, at 116–17. This Article calculates those rates using the online tool suggested by PCAST. *See* PCAST REPORT, *supra* note 2, at 153.

136. Law & Morris, *supra* note 122, at 1706. Such an approach is necessary to account for the reality that class eliminations neither so much as require firearms examiners to perform the totality of their method, *see supra* Section II, nor pose anything resembling the level of difficulty inherent to the subjective comparison of individual characteristics, *cf.* Keisler, *supra* note 128, at 56, 58 (describing how examiners achieved 100 percent specificity on out of class comparisons versus only 73.1 percent specificity on within-class comparisons). It is also consistent with the way scientists have distinguished between, and evaluated separately, performance on single-source DNA versus complex mixtures of DNA. *See* PCAST REPORT, *supra* note 2, at 7–8; Butler, *supra* note 110, at 3.

137. Guyll et al., *supra* note 68, at 9; Baldwin et al., *supra* note 49, at 3; Keisler et al., *supra* note 128, at 56–57.

138. Best & Gardner, *supra* note 15, at 31; Monson et al., *supra* note 68, at 88.

purposeful approaches designed to guarantee the inclusion of challenging comparisons, they have produced false positive estimates that simply do not square with claims like a "practical impossibility" of error.[139] With that brief outline alone, this article has backed up its claims both that courts citing false positive rates of 2 percent or less have necessarily ignored studies to the contrary, as well as that researchers have made misidentifications appear rare only by avoiding difficult comparisons. Nonetheless, it bears digging just a bit deeper to highlight the immense challenges confronting firearms examiners (and likely to produce error) and to tease out the relationship between accuracy rates and difficulty-levels; in other words, to underscore that the higher false positive rates generated to date for firearms examination better estimate the extent of potentially deficient performance in the real-world of casework.

	AFTE FPR	PCAST FPR	DIFFICULTY
Baldwin *et al.* (2023)[a]	1.1% (.6%, 1.5%)	1.5% (1%, 2.3%)	E
Keisler *et al.* (2018)[ac]	0% (0%, .5%)	0% (0%, .7%)	E
Guyll *et al.* (2023)[a]	.6% (0%, 1.3%)[d]	.9% (0%, 2.1%)[d]	E
Best & Gardner (2022)[bc]	.6% (.04%, 3.1%)	4.2% (.11%, 21.1%)	M
Monson *et al.* (2023)[ab]	.8% (.6%, 1.1%)	1.9% (1.4%, 2.6%)	M
Pauw-Vugts *et al.* (2013) [ab]	6.4% (3.9%, 9.7%)	9.5% (5.9%, 14.3%)	H
Mattijssen *et al.* (2020)[a]	5.6% (4.6%, 6.8%)	10.8% (8.8%, 13.1%)	H

Table 1: False Positive Rates in Pairwise Firearms Examination Validation Studies[140]

A. FAR FROM A WALK IN THE PARK: THE CHALLENGES CONFRONTING FIREARMS EXAMINERS

Initially, it would be deeply unfair to criticize firearms examiners for dodging difficulty in their accuracy studies if, in fact, the practice of comparing fired bullets and cartridge cases is inherently and universally simplistic and straightforward. But the existence of challenging pairs of different source bullets

139. Pauw-Vugts et al., *supra* note 15, at 117; Mattijssen et al., *supra* note 15, at 5–6.

140. The AFTE false positive rate includes inconclusives in the denominator. The PCAST false positive rate excludes inconclusives entirely. Clopper-Pearson Exact 95% confidence interval lower and upper bounds provided in parentheticals. Difficulty level assigned as follows: E=easy, no efforts to include coincidentally similar pairs; M=medium, coincidentally similar pairs assumed due to the use of consecutively manufactured samples; H=hard, purposeful / conscious attempts to seek out and include coincidentally similar pairs.

[a] Denotes cartridge case comparisons.

[b] Denotes bullet comparisons.

[c] Denotes calculations exclude out-of-class comparisons.

[d] Denotes confidence level figures taken directly from study and not independently calculated.

and cartridge cases (of close non-matches bearing substantial coincidental similarities) falls beyond reasonable debate. Since as far back as the 1950s, examiners have known that bullets fired from different guns (including even those manufactured in separate runs) can display levels of similarity exceeding, not just some matching pairs, but the average agreement seen in bullets fired by the same gun.[141] And for nearly as long, the field has rung its hands over the existence of subclass characteristics, and the incredible difficulty of distinguishing them from the kinds of "individual" markings more suitable to source attribution,[142] with one examiner going so far as to refer to them as a "specter" that has "has loomed over the field of firearms identification for a number of years."[143] Such characteristics can produce truly striking levels of agreement on bullets and cartridge cases fired by *different* guns, enough to cause examiners to wonder whether the AFTE *Theory of Identification* itself "may need to be reconsidered,"[144] and to opine that, at least in certain cases, "a correct identification of the firearm on the basis of the breech face and firing pin impression respectively, [will] . . . be hardly possible."[145] They provoked a rash of misidentifications in the 1980s.[146] And, though discovered primarily (if not exclusively) through the workings of fate rather than systemic searches,[147] have been documented on dozens of occasions across essentially every ammunition surface relied on by examiners;[148] one critic with expertise in manufacturing has

141. *See* Alfred A. Biasotti, "*A Statistical Study of the Individual Characteristics of Fired Bullets*, 4 J. Forensic Scis. 34, 38–40 (1959); *cf.* Jerry Miller & Michael McLean, *Criteria for Identification of Toolmarks*, 30 AFTE J. 15 (1998); Jerry Miller, *Criteria for Identification of Toolmarks Part II* Single Land Impression Comparisons*, 32 AFTE J. 116 (2000). Critics of the field have summarized these findings as follows: "Among those publications that hint at the nature and scope of the problem, one found up to 52% matching lines in a known non-match and another only 21-24% (steel-jacketed bullets) and 36-38% (non-jacketed bullets) concordance on bullets fired from the same gun. It has been observed that there are typically 2 and 3 times more matching striations in known non-matches (fired in different guns) than in those fired in the same gun." Tobin & Blau, *supra* note 51, at 136.

142. *See* David Q. Burd & Allan E. Gilmore, *Individual and Class Characteristics of Tools*, 13 J. FORENSIC SCIS. 390 (1968).

143. Gene C. Rivera, *Subclass Characteristics in Smith & Wesson SW40VE Sigma Pistols*, 39 AFTE J. 247 (2007). Other firearms examiners have expressed similarly, strongly-worded worry about such characteristics. *See, e.g.*, Fabiano Riva, *Objective Evaluation of Subclass Characteristics on Breech Face Marks*, 62 J. FORENSIC SCIS. 417 (2017) ("recognizing subclass characteristics is not an easy task, and some have rightly indicated that the ability of examiners to detect them is not well established"); Ronald G. Nichols, *Defending the Scientific Foundations of the Firearms & Toolmark Identification Discipline: Responding to Recent Challenges*, 52 J. FORENSIC SCIS. 586, 587 (2007) ("[T]he difficulty of addressing subclass characteristics is not in debate.").

144. Rivera, *supra* note 143, at 250 (saying of markings on the breech faces of two different Smith & Wesson firearms: "it is hard to imagine any better agreement between these two tools").

145. M. S. Bonfanti & J Dekinder, *The Influence of Manufacturing Processes on the Identification of Bullets & Cartridge Cases – A Review of the Literature*, 39 SCI. & JUST. 3, 5 (1999).

146. *See* Bruce Moran, *A Report on the AFTE Theory of Identification and Range of Conclusions for Tool Mark Identification and Resulting Approaches to Casework*, 34 AFTE J. 227 (2002).

147. *See, e.g.*, Rivera, *supra* note 143, at 247–49.

148. *See, e.g.*, M. Bar-Adon, L. Bokobza, Asaf Hazon & R. Siso, *Subclass Characteristics Found on Tactical-Hulk Semi-Automatic Pistols*, 50 AFTE J. 38, 39 (2018); Steve Kramer, *Subclass Characteristics on Firing Pins Manufactured by 'Metal Injection Molding'*, 44 AFTE J. 364, 365 (2012); William Matty & Torrey

even "observed that the majority of manufacturing marks . . . imparted to work pieces are subclass in nature."[149] Yet, despite all that, in the decades since discovering subclass characteristics, the field of firearms examination has not developed rules for distinguishing them from individual characteristics.[150] Indeed, what guidance leaders in the field have offered (for example that the land impressions of bullets or the aperture shear of cartridge cases will reliably be free from subclass influence) has not stood the test of time.[151] As manufacturers evolve and refine their means of production, the risk of misidentification posed by subclass characteristics will only increase.[152] But absent further development of firearms examination methods, practitioners will meet this threatening tide armed with nothing but "training and experience"[153] (a grossly inadequate response as we will shortly see).[154]

Johnson, *A Comparison of Manufacturing Marks on Smith & Wesson Firing Pins*, 16 AFTE J. 51 (1984); Evan Thompson, *False Breech Face ID's*, 28 AFTE J. 95 (1996); Michael Lee, *Subclass Carryover in Smith & Wesson M&P 15-22 Rifle Firing Pins*, 48 AFTE J. 27, 29 (2016); Vyacheslav Polosin, *Subclass Characteristics in Extractor Groove of Winchester Cartridges*, 48 AFTE J. 50 (2016); Alicia K. Welch, *Breech Face Subclass Characteristics of the Jiminez JA Nine Pistol*, 45 AFTE J. 336, 343 (2013); Frederic A. Tulleners & James S. Hamiel, *Subclass Characteristics of Sequentially Rifled 38 Special S & W Revolver Barrels*, 31 AFTE J. 117 (1999); Ronald Nies, *Anvil Marks of the Ruger MKII Target Pistol-An Example of Subclass Characteristics*, 35 AFTE J 75 (2003); Patrick D. Ball, *Toolmarks Which May Lead to False Conclusions*, 32 AFTE J. 292 (2000); Susan M. Komar & Gregory E. Scala, *Examiners Beware New Bolt Cutter Blades-Class or Individual*, 25 AFTE J. 298 (1993); Salvatore LaCova, *et al*, *Subclass Characteristics on CCI Speer Cartridge Case Heads*, 42 AFTE J. 281 (2010); Laura L. Lopez & Sally Grew, *Consecutively Machined Ruger Bolt Faces*, 32 AFTE J. 19 (2000); E. J. A. T. Mattijssen & Wim Kerkhoff, *Subclass Characteristics in a Gamo Air Rifle Barrel*, 45 AFTE J. 281 (2013); Peter Lardizabal, *Cartridge Case Study of the Heckler & Koch USP*, 27 AFTE J. 49 (1995); Tsuneo Uchiyama, *Similarity Among Breech Face Marks Fired from Guns With Close Serial Numbers*, 18 AFTE J. 15 (1986).

149. Spiegelman & Tobin, *supra* note 59, at 128.

150. *See* Biasotti & Murdock, *supra* note 31, at 18–19 ("Because what would constitute these subclass features is a function of the relative hardness of the tool, the material, and the dynamics of the cutting process, it is not currently possible to describe them in quantitative terms."); Schwartz, *supra* note 31, at 9 (noting that firearms examiners have no rules or statistics for the frequency of subclass marks, how they can be identified, or how long they may last, so that "examiners can only rely on their personal familiarity with types of forming and finishing processes and their reflections in toolmarks."); *Monteiro*, 407 F. Supp. 2d at 371 ("[O]ne critical problem with the AFTE Theory is the lack of objective standards for deciding whether a particular mark is a subclass or individual characteristic. . . . [T]here is no generally accepted standard for distinguishing between class, subclass, and individual characteristics.").

151. *See, e.g.*, Ronald Nichols, *Subclass Characteristics: From Origin to Evaluation*, 50 AFTE J. 68, 83 (2018).

152. *See* F. H. Cassidy, *Examination of Toolmarks from Sequentially Manufactured Tongue-and-Groove Pliers*, 25 J. FORENSIC SCIS. 796, 797 (1980) ("Modern mass-production methods for tools dictate the necessity of minimizing the manufacturing steps in order to make tool production as economical as possible. When this occurs, the manufacturing process could turn out consecutively manufactured parts that would have similar surface conditions."); Bonfanti & Dekinder, *supra* note 145, at 4 ("[A]s the techniques of firearms manufacture have evolved, following mostly commercial rather than forensic arguments, [their foundational assumptions] need to be verified on a regular basis.").

153. *See, e.g.*, United States v. Shipp, 422 F. Supp. 3d 762, 781 (E.D.N.Y. 2019) (criticizing firearms examination because "[t]he determination that the similarly between two sets of toolmarks indicates sufficient agreement between them and is not, instead, a result of subclass characteristics or random similarities between different firearms is left to the examiner's training and experience").

154. *See infra* Part IV.C.

What's more, we can expect technology to drive up the difficulty of firearms comparisons not just by increasing the prevalence of subclass characteristics, but also by forcing examiners to confront more cases originating from database leads. Specifically, the National Integrated Ballistic Information Network (NIBIN) allows law enforcement to enter images of cartridge cases it has recovered for automated comparison to the several million others it stores.[155] If the system returns a "lead" a firearms examiner may the be asked to "confirm" that the cartridge cases in question match (*i.e.*, were fired by the same gun).[156] This technology has generated hundreds of thousands of leads across its twenty-five-year history, and its use appears to be on the rise (of the 640,000 total leads issued by the system, 189,000 were from 2022 alone).[157] But problematically, as reliance on databases like NIBIN increases, so too will the rate at which examiners must confront close non-matches with the effect that "[e]rror rates (especially of the false-positive type) may increase."[158]Though concern and conversation about the ability of examiners to grapple with database close non-matches has primarily taken place in the realm of fingerprint comparisons,[159] there is no reason to consider firearms examination immune to an enhanced risk of misidentification stemming from NIBIN use.[160] Taken together with concerns related to coincidental similarity and subclass influence more generally, the prevalence of database leads firmly establishes the challenging nature of bullet and cartridge case comparisons, leaving as the only remaining question: Have firearms examiners truly explored the full range of difficulty we now know they will inevitably face in casework?

---

155. *See* Bureau of Alcohol, Tobacco, & Firearms, *Fact Sheet - National Integrated Ballistic Information Network*, https://www.atf.gov/resource-center/fact-sheet/fact-sheet-national-integrated-ballistic-information-network (last visited June 16, 2024).

156. *Id.*

157. *Id.*

158. Kellman et al., *supra* note 114, at 2; *see* Eldridge et al., *supra* note 116, at 2 ("Because of the nature of the algorithms used in Automated Fingerprint Identification Systems (AFISs) (which are designed to find the algorithmically closest matches contained in the database), combined with the ever-enlarging set of prints in their gallery, it is expected that examiners will increasingly face comparisons involving CNMs."); Koehler & Liu, *supra* note 116, at 130 ("The use of these databases, particularly large ones, may increase the risk of a false identification because they may contain hard-to-distinguish CNM prints. Concern about false identifications from database-derived CNMs is not merely theoretical. The Brandon Mayfield case . . . shows both that database CNMs exist and may fool even the best fingerprint examiners."); Itiel E. Dror & Jennifer L. Mnookin, *The Use of Technology in Human Expert Domains: Challenges and Risks Arising from the Use of Automated Fingerprint Identification Systems in Forensic Science*, 9 L., PROBABILITY & RISK 47, 58 (2010) ("Remember that the very aim of AFIS, its purpose, is to seek out the most similar prints in the database by testing and comparing each stored print against the latent exemplar from the crime scene. The whole point is to find whatever prints most resemble the latent according to the algorithms and search parameters provided, whether or not the actual source of the latent is even in the database at all. AFIS must, by design, increase the chances that the examiner will be presented with quite similar look-alike prints, as compared to those prints presented if suspects were identified through traditional investigative techniques rather than AFIS.").

159. *See supra* note 158.

160. *See* Joseph J. Masson, *Confidence Level Variations In Firearms Identifications Through Computerized Technology*, 29 AFTE J. 42 (1997).

B. THE CONSEQUENCES OF IGNORING SAMPLE DIFFICULTY OUTRIGHT

Researchers studying the accuracy of firearms examination have acknowledged much of the above, noting, for example, that "manufacturing processes associated with different firearm models . . . represent[t] a factor that could affect comparison difficulty and thereby affect examiners' ability to make correct decisions."[161] But they have also simultaneously (and thus quizzically) failed to vary the difficulty of their different source pairs. In fact, and as noted above, three pairwise studies to date have referenced no efforts whatsoever to include coincidentally similar samples, no efforts, in other words, to measure performance on the kinds of challenging samples described in the previous section of this article. One specifically avoided selecting firearms that according to "anecdotal suggestions . . . might be 'too hard.'"[162] Another utilized only Glock-type cartridge cases for its within-class comparisons[163] (even though such samples have been described variously as "ideal,"[164] "readily identifiable,"[165] and "a best case scenario"[166] for practitioners), resulting in a test that took participants, on average, only seven minutes per comparison and was, even by the admission of practitioners, "easy."[167] And the last selected two firearms models with an eye to varying the quantity of information available to examiners rather the extent of coincidental similarity, predicably affecting the percentage of inconclusive responses without any impact on the false positive rate. In fact, the model those authors described as "more difficult" actually provoked fewer misidentifications (by number and percentage).[168]

But all that said, the consequences of failing to seek out and include close non-matches to the value of false positive rates generated by these studies only really come fully into view by comparison to better-designed research efforts.

---

161. Guyll et al., *supra* note 68, at 4.

162. Baldwin et al., *A Study of Examiner Accuracy in Cartridge Case Comparisons. Part 1: Examiner Error Rates*, supra note 49, at 2.

163. *See* Keisler et al., *supra* note 128, at 56–57.

164. James E. Hamby, Stephen Norris & Nicholas D. K. Petraco, *Evaluation of GLOCK 9mm Firing Pin Aperture Shear Mark Individuality Based On 1,632 Different Pistols by Traditional Pattern Matching and IBIS Pattern Recognition*, 61 J. FORENSIC SCIS. 170, 172 (2016) ("GLOCKs are ideal in the sense that they are very well known to generally produce well-defined firing pin aperture shear on the primer of cartridge cases fired from them. Thus, a false match rate estimate on these provides a 'baseline' lower bound on the false match rate for more difficult toolmark comparisons.").

165. Stephen G. Bunch & Douglas P. Murphy, *A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases*, 35 AFTE J. 201, 202 (2003) ("[M]arks imparted to cartridge case primers from Glock breechfaces are, under normal circumstances, readily identifiable.").

166. Stroman, *supra* note 64, at 171 ("A shooting case involving Glock-fired cartridge cases generally represents a best case scenario for any firearms examiner because of the relative ease with which the firing pin aperture shear marks on these cartridge cases can be identified"); *see also* Baldwin et al., *supra* note 49, at 2 (reporting examiner suggestions that Glocks might be "too easy" for use in validation testing).

167. *See* Keisler et al., *supra* note 128, at 57 (reporting participants, on average, took 2.35 hours to complete the study's 20 comparisons); Trans. of Proceedings, Illinois v. Winfield, 15CR14066-01, at 196–97 (Cir. Ct. Cook Cnty. Mar. 1, 2022) (Todd Weller, a firearms examiner conceded that Keisler et al., *supra* note 128, should be considered "at least a somewhat easy test").

168. *See* Guyll et al., *supra* note 68, at 4, 7–8.

In other words, even when it comes to guns that, at a general level, provide examiners with the most simplistic eliminations, we should not so readily assume that "low" false positive rates generated by studies with minimal sampling (just eight guns total)[169] tell the whole story. If that were the case then the results of Mattijssen et al., which recorded a false positive rate potentially as high as 13.1 percent looking only at photos of Glock cartridge cases,[170] would not have been possible. These contrary results, however, are straightforward to reconcile. While Keisler et al. looked only at a small number of guns and selected them at random with no attention to the issue of coincidental similarity, Mattijssen et al. not only encompassed a far larger number of Glock firearms (two hundred), it also included comparisons specifically because they had challenged (or provoked error by) an algorithmic comparison tool.[171] Thus, while Keisler et al. perhaps should not be faulted generally for exploring performance on Glocks (they are, after all, prevalent amongst guns recovered in relation to criminal activity),[172] their failure to account for spectrum bias is both inexcusable and deeply misleading. If even slightly more expansive sampling and minimal attention to the inclusion of difficult pairs (like that seen in Mattijssen et al.) can so handily unseat firearms examination from even its preferred high ground of "readily identifiable" comparisons, then we must conclude that law enforcement claims to misidentification rates below 2 percent in casework are creatures not of science, but of fiction.

C. THE INADEQUACY OF CONSECUTIVE MANUFACTURE STUDIES

Firearms examiners would likely respond to the above by pointing to studies utilizing consecutively manufactured guns, which they have described as presenting "[t]he worst case scenario" for practitioners."[173] But while studies that avoided challenging comparisons outright have, predictably, generated misleading low false positive rates, those that utilized that modest effort of including consecutively manufactured samples to inject difficulty scarcely fare better. As much follows from the reality that the largest example of the latter (Monson et al.) produced estimates of the potential for misidentification no larger than the former (Baldwin et al.), despite allegedly including more

169. *See* Keisler et al., *supra* note 128, at 56.

170. *See* Mattijssen, *supra* note 15, at 7 (reporting a 10.8 percent false positive rate when excluding inconclusives); *supra* Table 1.

171. *See* Mattijssen, *supra* note 15, at 2, 5–6.

172. *See* BATF: VOLUME II, PART III *supra* note 126, at 19 (noting that Glocks account for close to 20 percent of the crime-involved handguns traced by the ATF between 2017 and 2021); Fransisco Alvarado, *Glock Pistols Are the Overlooked Weapon in American Mass Shootings*, VICE NEWS (June 21, 2016), https://www.vice.com/en/article/gy9nj4/glock-pistol-omar-mateen-orlando-mass-shooting (tracing the involvement of Glock firearms across multiple mass shootings).

173. Best & Gardner, *supra* note 15, at 28.

challenging comparisons.[174] But if that doesn't suffice to persuade, criticisms of a narrow focus on consecutively manufactured guns alone abound.[175] Such criticisms suggest that said design may actually give examiners advantages unavailable in casework,[176] and transcend the theoretical.[177] In fact, independent researchers reanalyzing the raw response data from Monson et al. concluded that that study's claim to have included challenging comparisons "should not be accepted" because, though purportedly focused only on samples "likely to be highly similar,"[178] participants based 449 of their cartridge case eliminations (32.1 percent) on obvious differences in class characteristics.[179] According to those critics, their findings, "undermin[e] the reported false positive error rates . . . [as] artificially low because there is virtually no risk of a different- class elimination being called an identification."[180]

More to the point, these studies have failed to assess the very reason for their underlying assumption that consecutively manufactured samples will produce more challenging or coincidentally similar pairs, namely, sub-class characteristics.[181] Rather than in any way confirm the presence of such markings amongst their samples, both studies instead relied on a wing and a prayer, hoping that because the firearm models they selected "anecdotally" have a tendency to

174. *See* Monson et al., *supra* note 68, at 94 (claiming that "Experimental parameters of the present study were challenging by design"). The other generated a larger false positive rate only tangentially. Like Guyll et al., *supra* note 68, it took measures (including damaged bullets) to impact the quantity of information available to participants. *See supra* Part IV.B; Best & Gardner, *supra* note 15, at 32. Those measures seemingly worked to great effect, diminishing specificity to an absurdly low 13.1 percent. *See supra* note 68. Only because of that paucity of within-class eliminations did the solitary false positive in said study produce such an incredibly wide confidence interval range of less than 1 percent all the way to 21.1 percent. *Id.*

175. *See* BALLISTIC IMAGING, *supra* note 33, at 70–72 (attacking earlier studies of consecutively manufactured guns for their small samples sizes and failure to consider whether sequential serial numbers actually indicate consecutive manufacture); NAS REPORT, *supra* note 2, at 155 (criticizing consecutive manufacture studies for "a heavy reliance on the subjective findings of examiners rather than on the rigorous quantification and analysis of sources of variability."); Biasotti & Murdock, *supra* note 33, at 19 ("The information gained from such studies is therefore only of value to the examiner who conducted the study; or to the examiners trained or supervised by that examiner."); Tobin & Blau, *supra* note 51, at 139 ("As it turns out, careful analysis for both internal and external validity of the various putative validation studies that currently exist reveals them to be nothing more than very limited proficiency tests of the participating examiners . . .in addition to the fact that they do not circumstantially mirror casework."); Mark Page et al., *supra* note 90, at 15 (noting that even legitimate studies of consecutively manufactured guns fail entirely to address the issue of random matching and examiner performance on closest potential non-matches).

176. *See* Dorfman & Valiant, *supra* note 15, at 5 ("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.").

177. *See* Nicholas Scurich & Hal Stern, *Commentary on: Monson KL, Smith ED, Peters EM. Accuracy of Comparison Decisions by Forensic Firearms Examiners*, 68 J. FORENSIC SCIS. 1093 (2023).

178. Monson et al., *supra* note 68, at 87.

179. Scurich & Stern, *supra* note 177, at 1094. This phenomenon was less pronounced in, but still applicable to, bullet comparisons. *See* Keith L. Monson, Erich Smith & Eugene M. Peters, *Authors' response to Scurich et al Commentary on: Monson KL, Smith ED, Peters EM. Accuracy of comparison decisions by forensic firearms examiners*, 68 J. FORENSIC SCIS. 1095 (2023).

180. Scurich & Stern, *supra* note 177, at 1094.

181. *See* Best & Gardner, *supra* note 15, at 28; Monson et al., *supra* note 68, at 94.

display sub-class characteristics, their limited sampling of just single manufacturing runs would follow suit.[182] Some firearms examiners have, if unintentionally, even admitted the inadequacy of such an approach, conceding (despite the existence of two pairwise, consecutive manufacture studies) that they "are unaware of any study that assess[es] the overall firearm and toolmark discipline's ability to . . . identify subclass marks."[183] But as was the case with studies involving Glocks, a comparison to a European research effort that went above and beyond such dangerous assumptions (went beyond conflating the *potential* for subclass characteristics with definitive presence amongst samples) should erase any remaining doubt about whether "low" false positive rates truly account for difficult, close non-match situations. That study, in reaction to proficiency exams that were "insufficiently challenging to be of use in demonstrating competence in comparison microscopy skills," set out to design a test with "realistic and challenging comparisons."[184] It did so in part by including cartridge cases *known* to display subclass characteristics on their breechfaces, and (even though images of precisely those casings had appeared in AFTE's journal) unsurprisingly recorded what its authors called a "disturbingly high" false positive rate overall (one potentially as high as 14.3 percent).[185] Worse, performance on the samples that displayed those subclass characteristics barely (if at all) exceeded chance: participants committed misidentifications on 38.9 percent of their conclusive decisions (with an upper bound of 56.5 percent).[186] The data do not lie: until examiners prove otherwise there is little reason to assume consecutive manufacture studies suffice to cover the full range of difficulty in casework, and significant hints that the false positive rates they have generated are inconsistent with (and misleadingly rosier than) the performance we can expect on challenging cases foreseeable in the real world.

***

That courts, by and large and despite all the above, have been unconcerned about the potential for spectrum bias and taken in by law enforcement claims regarding firearms examination's allegedly minimal proclivity for misidentifications is, in some sense, not surprising. After all, firearms examiners have gone to great lengths to exile evidence to the contrary beyond the reaches of judicial scrutiny: repeatedly eliding mention of studies with double-digit false

182. *See id.*

183. Org. of Sci. Area Comms. for Forensic Sci. Firearms & Toolmarks Subcommittee, *Research Needs Assessment Form: Assessment of Examiners' Toolmark Categorization Accuracy* (2021), https://www.nist.gov/organization-scientific-area-committees-forensic-science/osac-research-and-development-needs.

184. Pauw-Vugts et al., note 15, at 117.

185. *See id.* at 124–26.

186. *See id.* at 125.

positive rates in summaries of foundational research and in sworn testimony,[187] as well as advancing empirically unsupported and unsupportable descriptions of their favored studies as "challenging."[188] But judicial willingness to accept this dominant narrative at face value must end. Challenging samples with substantial levels of coincidental similarity (close non-matches) have and will come across the desks of examiners in casework. The leap of faith firearms examiners have demanded in relation to the representativeness of their error rate studies and their coverage of the full range of casework difficulty, falters both in the shallows ("easy" comparisons like Glocks) and the depths (difficult comparisons requiring examiners to rule out subclass agreement). And those studies showing double-digit false positive rates cannot reasonably be discounted or distinguished (as some have already tried with regards to the European examples)[189] based on the conclusion scales they employed,[190] the national origin of participants,[191] or their use of photos and casts as opposed to original, physical cartridge cases.[192] Indeed, law enforcement efforts to exclude these studies cannot be taken as sincere given that the FBI, for example, ultimately cites to other studies (just so happening to have generated false positive estimates it preferred) that involved European examiners, differing conclusion scales, and cast samples, as well as one still pending publication.[193] Ultimately

187. *See* FBI Statement, *supra* note 65, at 18–20 (mentioning none of Mattijssen et al., *supra* note 15; Pauw-Vugts et al., *supra* note 15; or Knapp & Garvin, *supra* note 15); DOJ Statement, *supra* note 65, at 23–24 (same); Org. of Sci. Area Comms. Firearm & Toolmark Subcommittee, *Response to the President's Council of Advisors on Science and Technology (PCAST)*, (Dec. 23, 2015), https://www.nist.gov/organization-scientific-area-committees-forensic-science/firearms-toolmarks-subcommittee (mentioning neither Pauw-Vugts et al., *supra* note 15; nor Knapp & Garvin, *supra* note 15); Trans. of Proceedings, Illinois v. Winfield, 15CR14066-01, at 242–43 (Cir. Ct. Cook Cnty. Mar. 1, 2022). Firearms examiner Todd Weller admitted he had failed to mention existing studies with higher false positive rates multiple times during sworn testimony. *Id.*

188. *See* Monson et al., *supra* note 68, at 94.

189. *See* United States v. Pete, No. 3:22CR48-TKW, 2023 WL 4928523, at *4 (N.D. Fla. July 21, 2023) (discounting Pauw-Vugts et al., *supra* note 15); FBI Statement, *supra* note 65, at n.13.

190. In one, 75 percent (58 of 77) participating examiners used a categorical conclusion scale (identification, inconclusive, elimination). *See* Mattijssen, *supra* note 15, at 4. And the strength-of-support scale used in the other is nearly identical to the AFTE range: two of the conclusions (identification and elimination) are identical, *see* Pauw-Vugts et al., *supra* note 15, at 120, and even the FBI has acknowledged the insignificance of any differences between the two, calling them "highly comparable" and distinguished only by "nomenclature." Monson et al., *supra* note 68, at 95 ("Alternative scales, which describe conclusions in terms of strength of support, are under consideration. If adopted by the community, the value of studies using the AFTE range will endure. The proposed scale is highly comparable to the AFTE range, being essentially a change in nomenclature. The term Elimination is replaced by Exclusion, while Identification remains. The middle three conclusions of the proposed scale closely approximate the definitions of the three AFTE levels of Inconclusive.").

191. American examiners and laboratories participated in both studies, with neither suggesting their performance exceeded that of European examiners. *See* Mattijssen, *supra* note 15, at 4; Pauw-Vugts et al., *supra* note 15, at 116.

192. *See* Pauw-Vugts et al., *supra* note 15, at 126 (noting that "while the quality of the material can cause a 'B' or 'C' conclusion [i.e. inconclusive responses] instead of an 'A' [i.e. an identification] it would not cause a false exclusion or vice versa").

193. *See* FBI Statement, *supra* note 65, at 18–20 (citing Law & Morris, *supra* note 122); Bajic et al., *supra* note 85; W. Kerkhoff, R. D. Stoel, C. E. H. Berger, E. J. A. T. Mattijssen, R. Hermsen, N. Smits & H. J. J. Hardy, *Design & Results of an Exploratory Double-Blind Testing Program in Firearms Examination*, 55 SCI. & JUST. 514 (2015).

then, to the extent we know anything about the false positive rate firearms examiners will bring to the momentous task of proving guilt and protecting innocence (and in truth we know far too little),[194] all signs point to a field that cannot be trusted, to a field whose practitioners (when we account for spectrum bias) are not likely to misidentify evidence in only 2 percent or less of cases, but instead in double-digit percents of cases overall and at rates worse than a flip of a coin when confronting sub-class characteristics (*i.e.*, worse than random chance in *potentially* and *unpredictably* any given case they encounter).[195]

CONCLUSION: STEPPING OFF THE PENROSE STAIRS

As this Article has demonstrated, courts that have parroted law enforcement by accepting and repeating proffered false positive rates for firearms examination of 2 percent or less have wildly underestimated the field's potential to misidentify the source of fired bullets and cartridge cases, and have thereby imperiled the innocent. Such figures stand, necessarily and only, by sidestepping the sweeping criticisms by scholars on the trustworthiness of existing error estimates (merely outlined in this paper), dispensing with nuanced analysis of whether the accuracy studies from which they emerge adequately explored the full range of difficulty expected in casework, and turning a blind eye to those research efforts which have measured far higher rates soaring into the double digits and (at least on certain samples) approaching mere chance. Indeed, as regards the last of those, courts comfortable with their admissibility decisions hinging on a misidentification rate of under 2 percent have, even if unintentionally, bypassed nearly half the story: Three of seven pairwise studies[196] of firearms examiner accuracy have produced false positive rates with

194. *See supra* Part II.

195. Though beyond the scope of this article, it does bear mentioning that some courts have expressed a belief that verification will further reduce false positive rates in casework. *See, e.g.*, United States v. Rhodes, No. 3:19-CR-00333-MC, 2023 WL 196174, at *4 (D. Or. Jan. 17, 2023); *but see* Abruquah v. Maryland, 296 A.3d 961, at 687 (2023) ("[T]he record also contains evidence that severely undermines the value of some of those same standards and controls. For example, one control touted by advocates of firearms identification is a requirement that a second reviewer confirm every identification classification . . . [but] the confirmatory review process is not blind, meaning that the second reviewer knows the conclusion reached by the first. Even more significantly, Dr. Hamby testified that in his decades of experience in firearms identification in multiple laboratories in multiple states, he was not aware of a single occasion in which a second reviewer had reached a different conclusion than the first."). Such an assumption, however, is unacceptable without empirical data. *See* PCAST REPORT, *supra* note 2, at 96 ("[I]t would not be appropriate simply to infer the impact of independent verification based on the theoretical assumption that examiners' errors are uncorrelated."). That is especially true because non-blind verification in firearms examination seems to produce few disagreements. *See* Erwin J.A.T. Mattijssen, Cilia L. M. Witteman, Charles E. H. Berger & Reinoud D. Stoel, *Cognitive Biases in the Peer Review of Bullet and Cartridge Case Comparison Casework: A Field Study*, 60 SCI. & JUST. 337 (2020). And at least one study has produced a false positive despite allowing for verification to occur. *See* Best & Garnder, *supra* note 15.

196. *See supra* Table 1.

upper bounds exceeding even PCAST's generous "threshold" for validity of 5 percent.[197]

Worse, change on the part of the field of firearms examination seems unlikely absent the precursor of more skeptical judicial review.[198] Far from paranoia, that admittedly pessimistic prediction already finds support in the approach of studies involving the emerging technology of virtual comparison microscopy (tools which visualize, for comparison by human examiners, the surfaces of bullets and cartridge cases using 3D topography scanning and measurements).[199] While researchers have readily acknowledged the need to validate such technology, including specifically by addressing the impact (if any) that a shift from conventional methods to VCM might have on practitioner error rates,[200] they have also, and unfortunately, carried forward all the same signs of and capitulation to spectrum bias that pervade studies of traditional comparison methods.[201] Of the three projects to have tackled the issue, one (already quoted above) paradoxically claims to have selected comparison items "to represent the type and variety of exhibits that would be expected to be encountered in day-to-day casework," despite blustering that "[t]here was no intention to select the pairs in the true elimination sets that would attempt to lead the participants into making a false positive source-attribution conclusion (e.g. strong carry-over of subclass characteristics between the pairs)." In fact, the latter appears not even to have given the authors pause in reporting and discussing the significance of false positive and specificity performance.[202] Another utilized elimination comparison sets so distinctly dissimilar, so

197. *See* PCAST REPORT, *supra* note 2, at 151–52. The PCAST Report does not specify whether its 5 percent figure applies to point estimates of a method's false positive rate or to confidence interval upper bounds (nor does it specify whether said figure should turn only on conclusive decisions), but given the report's focus on each of the latter, it seems safe to assume it meant that the upper bound for a measured false positive rate among conclusive decisions should not exceed 5% or be "considerably lower." *See id.* at 152–53.

198. *See, e.g.*, D. Michael Risinger & Michael J. Saks, *A House with No Foundation*, 20 ISSUES SCI. & TECH. 35 (2003) (explaining that, bolstered by judicial decisions admitting the testimony of practitioners without conducting searching inquiries or demanding foundational validity, forensic communities have dismissed research that might uncover limitations as a "net loss").

199. *See* Chapnick et al., *supra* note 122, 557–58 (2021); Pierre Duez, Todd Weller, Marcus Brubaker, Richard E. Hockensmith & Ryan Lilien, *Development and Validation of a Virtual Examination Tool for Firearm Forensics*, 63 J. FORENSIC SCIS. 1069, 1069–70 (2018). These tools, new though they might be, have already begun to make their way into the courts. *See* New Jersey v. Ghigliotty, 232 A.3d 468, 485 (App. Div. 2020).

200. *See* Chapnick et al., *supra* note 122, at 558 ("[I]ncorporation of new technology into a laboratory requires validation and establishment of error rates. It is only by establishing well-founded error rates that the technology will truly benefit the criminal justice system."); Duez et al., *supra* note 199, at 1072 ("Prior to its routine use in a crime laboratory, it is important to validate the use of any new technology . . . it will be necessary to demonstrate that virtual microscopy can reliably achieve comparison results at least as good as that obtained using conventional methods.").

201. This has not stopped courts from quizzically citing them in support of claims that traditional comparison methods enjoy low false positive rates. *See, e.g.*, United States v. Harris, 502 F. Supp. 3d 28, 37–38 (D.D.C. 2020). This criticism is also not meant to apply to researchers exploring objective, comparison algorithms who, as noted above, have been far more conservative in describing the significance of their existing record of testing. *See supra* note 35.

202. Knowles et al., *supra* note 49, at 517, 522.

obviously fired by different guns (even by the admission of the study's designers),[203] that *untrained lawyers* were able to complete the test without committing misidentifications.[204] And examiner annotations from the last, tired preambles to the contrary aside,[205] show that it too failed to include different-source comparisons with coincidental similarities, and thus any chance of provoking or adequately measuring false positives.[206] If those exploring even the technologies and methods of firearms examination's future remain resolute in their commitment to the research deficiencies of the present and past, then the need for judicial action, and for greater vigor in the role of gatekeeper, has clearly reached an inescapable apex.

Yet, despite all the negative prognoses laid out in this Article, there are reasons for hope in the turning of the tide. Litigation focused on firearms examination's troubling predilection for misidentification may well rattle judicial complacency. After all, at least one judge, when confronted with some of the studies discussed herein showing higher false positive rates, likened the experience, en route to excluding cartridge case and bullet comparison testimony outright, to being taken "into an even more terror-filled room of the State's haunted house of firearms identification evidence . . . [a] basement room of horrors."[207] The amendments to Rule 702, mentioned in this Article's Introduction, have already impacted judicial screening of firearms examination and been cited by the judge in *Briscoe* as part of her rationale for going further than circuit court precedent, casting doubt on error rate estimates, and precluding

203. *See* Duez et al., *supra* note 199, at 1080, 1083 (explaining that "it is difficult to infer the reason that a false identification was made[; . . . t]he shears are quite different").

204. *See* Illinois v. Winfield, No. 15CR14066-01, at 13 (Cir. Ct. Cook Cnty. Feb. 8, 2023) (on file with author); Scurich et al., *supra* note 26, at 5; Balko, *supra* note 49 (calling the situation "[o]ne almost comical example" of a pattern that "[p]ractitioner-administered tests also tend to be easier"). In fact, one of the study's authors, Todd Weller, actually conceded under oath that, although he had previously testified about its results and had not caveated for that earlier judge that they were based off simplistic comparisons, it would not "surprise" him if attorneys could complete the comparisons without committing a misidentification. Trans. of Proceedings, Illinois v. Winfield, 15CR14066-01, at 205, 211 (Cir. Ct. Cook Cnty. Mar. 1, 2022). That manner of describing the study had its intended effect; the judge in *Harris* saw the low false positive rate, but not the simplicity underlying it, and cited Duez et al., *supra* note 199, in support of his decision to admit firearms examination evidence. *See* 502 F. Supp. 3d at 37–38. The results from Winfield have since been replicated in a peer reviewed study involving 82 untrained attorneys whose "performance on different-source comparisons was essentially indistinguishable from that of trained examiners." Richard E. Gutierrez & Emily J. Prokesch, *The False Promise of Firearms Examination Validation Studies: Lay controls, simplistic comparisons, and the failure to soundly measure misidentification rates*, 69 J. FORENSIC SCIS 1334 (2024).

205. *See* Chapnick et al., *supra* note 122, at 568 ("The VCMER study involved 40 test sets covering a range of common firearm makes, models, and calibers. These sets include both well and minimally marked cartridge cases spanning a range of expected comparative complexity.").

206. *See id.* at 566–68; *Winfield*, No. 15CR14066-01, at 13. Similar to note 204 above, Todd Weller admitted in sworn testimony that participants repeatedly did not observe "features that are similar that could provoke a misidentification." Trans. of Proceedings, Illinois v. Winfield, 15CR14066-01, at 213–16 (Cir. Ct. Cook Cnty. Mar. 1, 2022).

207. *Winfield*, No. 15CR14066-01, at 24.

source attribution testimony by a firearms examiner.[208] And, of course, even acceptance of law enforcement claims regarding misidentification rates below 2 percent by judges has now repeatedly failed to translate into wholesale admissibility. In other words, judges have, and may continue to, accept that figure without correspondingly interpreting it as low enough to weigh in favor of admissibility.[209] But since these signs have not reached the status of norms, and since any corresponding hopes yet resemble ember more than flame, it bears lingering, before concluding entirely, on the insights of that last group of judges and on questions largely otherwise skirted by this article's focus on the defensibility of a 2 percent-or-less false positive figure: Should a misidentification in up to one in every fifty cases really qualify as rare enough to favor admissibility, and would the judges who have concluded as much have done so if the burden and consequences of error fell on their backs rather than those of defendants before their benches?[210]

Engaging these issues means reckoning with our criminal legal system's longstanding complicity in an "anti-Black punitive tradition" defined by "the habitual surveillance and incapacitation of racialized individuals and communities,"[211] as well as with the outsized role that forensic sciences have played in the current era of mass incarceration.[212] As Chris Fabricant, director of the Innocence Project's Strategic Litigation Department, has adroitly noted in pursuit of characterizing as "poor people's science" much of the forensic landscape: "That there are two systems of justice in America, one for the wealthy, one for the poor, is hardly a novel observation. But that there are two types of science, one for the rich and one for poor people, is less commonly understood."[213] On the first of those points, the statistics are staggering to say

208. *See* United States v. Briscoe, No. 20-CR-1777 MV, 2023 WL 8096886, at *4, 9, 12–13 (D.N.M. Nov. 21, 2023) (distinguishing United States v. Hunt, 63 F.4th 1229, 1240–41 (10th Cir. 2023) (affirming district court's decision to allow firearms examiner to opine on source with only limitation against absolute certainty claims)).

209. *See* United States v. Shipp, 422 F. Supp. 3d 762, 778 (E.D.N.Y. 2019); United States v. Adams, 444 F. Supp. 3d 1248, 1264 (D. Or. Mar. 16, 2020); Oregon v. Moore, No. 18CR77176, at 24–26 (Cir. Ct. Or. Aug. 8, 2023).

210. If nothing else, such a false positive rate exceeds, by a full order of magnitude, the one detected for latent print comparisons in the largest-ever accuracy study conducted for that field. *See* PCAST REPORT, *supra* note 2, at 98.

211. Elizabeth Hinton & DeAnza Cook, *The Mass Criminalization of Black Americans: A Historical Overview*, 4 ANN. REV. CRIMINOLOGY 261, 263 (2021).

212. *See, e.g.*, Sinha, *supra* note 63, 897 (2022).

> "Given that those targeted for prosecution and conviction are disproportionately Black, Brown, or otherwise of color, it comes as no surprise that those convicted by unreliable forensic evidence are also members of marginalized communities. The overlap between the increased use of forensic techniques and the mass expansion of the criminal legal system makes clear that those who have been hit hardest by nearly five decades of expanded criminalization, Black and Brown communities, are also the most likely to bear the brunt of flawed forensics in their cases." *Id.*

213. M. Chris Fabricant, *Poor People Science: Junk Science and the American Criminal Justice System*, MD. ST. BAR ASS'N (Mar. 16, 2023), https://www.msba.org/poor-people-science-junk-science-and-the-american-criminal-justice-system-2.

the least. Not only do black men comprise a disparate percentage of the prison population—13 percent of the United States population versus over 30 percent of those incarcerated[214]—they are also "seven times more likely than white Americans to be falsely convicted of crimes," and "are over-represented to a greater or lesser extent among exonerations for all major crime categories . . . except white collar crime."[215] And things are scarcely better for the indigent, who comprise around 80 percent of those targeted for prosecution.[216] But support for Mr. Fabricant's second contention about the disparate impact of forensic sciences is no less forthcoming. Despite the stock that jurors and judges place in forensic methodologies,[217] not to mention the latter's insistence that such techniques constitute engines of truth just as likely to exonerate the innocent as to convict the guilty,[218] faulty and misleading forensic evidence has wrought substantial harm in our system of criminal prosecutions, contributing to 25 percent of known exonerations overall and 50 percent of the smaller subset of exonerations uncovered through DNA testing.[219] As with mass incarceration more generally, race, disparity, and privilege all come into play: 53 percent of the 804 people, to date, declared

214. *See, e.g.*, Elizabeth Hinton, LaShae Henderson, & Cindy Reed, *An Unjust Burden: The Disparate Treatment of Black Americans in the Criminal Justice System*, VERA INST. OF JUST. (2018), www.vera.org/downloads/publications/for-the-record-unjust-burden-racial-disparities.pdf; Mike Wessler, *Updated Charts Provide Insights on Racial Disparities, Correctional Control, Jail Suicides, and More: New Data Visualizations Expose the Harms of Mass Incarceration*, PRISON POL'Y INITIATIVE (May 19, 2022), https://www.prisonpolicy.org/blog/2022/05/19/updated_charts.

215. Samuel L. Gross, Maurice Possley, Ken Otterbourg, Klara Stephens, Jessica Weinstock Paredes & Barbara O'Brien, *Race & Wrongful Convictions in the United States*, NAT'L REGISTRY OF EXONERATIONS, at 1 (2022), https://www.law.umich.edu/special/exoneration/Pages/about.aspx (further noting that black men comprise 53 percent, *i.e.*, four times their percentage of the population, of those convicted and later exonerated).

216. *See* Richard A. Oppe & Jugal K. Patel, *One Lawyer, 194 Felony Cases, and No Time*, N.Y. TIMES (Jan. 31, 2019), https://www.nytimes.com/interactive/2019/01/31/us/public-defender-case-loads.html; Mercedes Molina, *You Have the Right to Chronically-Underfunded and Overworked Counsel: The Need for Improved Support of Public Defense in Cook County and Beyond*, CHI. APPLESEED (Aug 11, 2023), https://www.chicagoappleseed.org/2021/08/11/your-right-to-chronically-underfunded-overworked-cook-county-public-defender; *see also* Bernadette Rabuy & Daniel Kopf, *Prisons of Poverty: Uncovering the Pre-Incarceration Incomes of the Imprisoned*, PRISON POL'Y INITIATIVE (July 9, 2015), https://www.prisonpolicy.org/reports/income.html ("[I]n 2014 dollars, incarcerated people had a median annual income of $19,185 prior to their incarceration, which is 41% less than non-incarcerated people of similar ages.").

217. *See generally* Jonothan J. Koehler, *Intuitive Error Rate Estimates for the Forensic Sciences*, 57 JURIMETRICS 152 (2017); Katie Kronick, *Forensic Science and the Judicial Conformity Problem*, 51 SETON HALL L. REV. 589 (2021).

218. *See, e.g.*, Illinois v. Jones, 2015 IL App (1st) 121016, ¶ 72 ("The reality in forensic science and its application to criminal cases and our justice system is that these human expert interpretations are highly probative and aid triers of fact and the police in not only convicting but also excluding suspects as perpetrators of crimes.").

219. *See % Exonerations by Contributing Factor*, NAT'L REGISTRY OF EXONERATIONS, https://www.law.umich.edu/special/exoneration/Pages/ExonerationsContribFactorsByCrime.aspx (last visited Nov. 30, 2023); Simon A. Cole, Vanessa Meterko, Sarah Chu, Glinda Cooper, Jessica Weinstock Paredes, Maurice Possley & Ken Otterbourg, *The Contribution of Forensic and Expert Evidence to DNA Exoneration Cases: An Interim Report*, NAT'L REGISTRY OF EXONERATIONS & INNOCENCE PROJECT (2022), https://www.law.umich.edu/special/exoneration/Pages/about.aspx.

innocent after convictions precipitated by faulty forensic testimony have been Black.[220]

Thus, the conclusion that our legal system abides such glaring inequity specifically because it is marginalized communities (the black, the brown, and the indigent of our society) that "bear the brunt of flawed forensics in their cases" is hard to escape.[221] In stark contrast to industries that more equally impact privileged segments of our population, forensic sciences have faced so little regulation that one scholar has quipped about "the paradoxical result that clinical laboratories must meet higher standards to be allowed to diagnose strep throat than forensic labs must meet to put a defendant on death row."[222] For decades, *Daubert* has worked miracles for the prosecution and for civil defendants (many of them corporations with Marianas-deep pockets) while hanging the indigent accused out to dry.[223] And the "carceral culture"[224] and prosecutorial-alignment of forensic scientists (a community, by the way, in which people of color are systemically underrepresented)[225] has "prevent[ed] the adversary process from working, as intended, to expose error."[226] Judges who have dismissed false positive rates of 2 percent or less as "low" act part and parcel with this vicious divide, because in other areas where the white and the wealthy of our society more frequently bear the risk of error, a failure rate of 2 percent would never suffice. Facing evidence that the Johnson & Johnson vaccine for COVID-19 had just a 0.00000323 percent chance of causing blood clotting (and an even smaller 0.00000048 percent chance of causing death), both the CDC and the FDA pulled back from their initial hopes in the utility of the vaccine, restricting access to only individuals who could not obtain alternative mRNA options.[227] And despite

220. *See Exonerations by State*, NAT'L REGISTRY OF EXONERATIONS, http://www.law.umich.edu/special/exoneration/Pages/Exonerations-in-the-United-States-Map.aspx (last visited June 16, 2024).

221. Sinha, *supra* note 63, at 897.

222. Paul C. Gianelli, *Crime Labs Need Improvement*, ISSUES SCI. & TECH. (2003), https://issues.org/giannelli.

223. *See, e.g.*, Erin Murphy, *Neuroscience and the Civil/Criminal* Daubert *Divide*, 85 FORDHAM L. REV. 619, 621–24 (2016); Peter J. Neufeld, *The (Near) Irrelevance of* Daubert *to Criminal Justice and Some Suggestions for Reform*, 95 AM. J. PUB. HEALTH S107, S110 (2005); Brandon L. Garrett & M. Chris Fabricant, *The Myth of the Reliability Test*, 86 FORDHAM L. REV. 1559 (2018); M. CHRIS FABRICANT, JUNK SCIENCE AND THE AMERICAN CRIMINAL JUSTICE SYSTEM 66–77 (2022). Scholars have even traced this issue in the specific context of firearms examination. *See* Garrett et al., *supra* note 1; Jim Hilbert, *The Disappointing History of Science in the Courtroom:* Frye, Daubert*, and the Ongoing Crisis of "Junk Science" in Criminal Trials*, 71 OKLA. L. REV. 759 (2019).

224. *See* Sinha, *supra* note 63, at 896–904.

225. *See* An-Di Yim, Jessica K. Juarez, Jesse R. Goliath & Isabel S. Melhado, *Diversity in Forensic Sciences: Black, Indigenous, and People of Color (BIPOC) Representation in Different Medicolegal Fields in the United States*, 5 FORENSIC SCI. INT'L: SYNERGY 100280 (2022).

226. Michael J. Saks, *Merlin and Solomon: Lessons from the Law's Formative Encounters with Forensic Identification Science*, 49 HASTINGS L.J. 1069, 1092 (1998).

227. *See* U.S. FOOD & DRUG ADMIN., CORONAVIRUS (COVID-19) UPDATE: FDA LIMITS USE OF JANSSEN COVID-19 VACCINE TO CERTAIN INDIVIDUALS (May 5, 2022), https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-limits-use-janssen-covid-19-vaccine-certain-individuals; CTRS. FOR DISEASE CONTROL & PREVENTION, INTERIM CLINICAL CONSIDERATIONS FOR USE OF COVID-19

the prevalence among Americans of aviophobia (the fear of flying),[228] airlines have been forced for decades to maintain an exemplary safety record, keeping accidents below 0.0000032 percent of departures.[229] Judges, privileged professionals that they are, must ask themselves why and how they can demand, why and how they can expect, such *de minimus* levels of risk for themselves and their families while accepting far less excellence from forensic industries bent on conviction. At bottom, no nation can survive—no ragged notion of justice can sustain itself—when the innocent-accused must face down a greater chance of catastrophe than travelers to a beach vacation. Barring firearms examination evidence will not alone remedy such glaring disparities between the privileged and the poor, between white and black, between the marginalized and the powerful, but that reality makes it no less necessary a step on the walk towards a safer and more equitable future.[230]

VACCINES: APPENDICES, REFERENCES, AND PREVIOUS UPDATES (Mar. 16, 2023), https://www.cdc.gov/vaccines/covid-19/clinical-considerations/interim-considerations-us-appendix.html#appendix-a.

228. *See* Daniel DeVise, *Up to 40 Percent of Americans Fear Flying. It's Easily Treated*, HILL (Mar. 6, 2023).

229. *See* INT'L CIVIL AVIATION ORG., SAFETY REPORT 5 (2021); INT'L CIVIL AVIATION ORG., SAFETY REPORT 13 (2017). Both of the cited reports (as well as several others) are available at https://www.icao.int/safety/pages/safety-report.aspx.

230. Simply calling out these disparities likely will not suffice to change hearts or minds. *See* Rebecca C. Hetey & Jennifer L. Eberhardt, *The Numbers Don't Speak for Themselves: Racial Disparities and the Persistence of Inequality in the Criminal Justice System*, 27 CURRENT DIRECTIONS PSYCH. SCI. 183 (2018) ("Ironically, exposure to extreme disparities can cause people to become more, not less, supportive of the very policies that create those disparities."). But hopefully, by "(a) offer[ing] context, (b) challeng[ing] associations, and (c) highlight[ing] institutions" specific to forensic evidence, this Article will have gone that necessary step further and forced, at least pause, into the otherwise reflexive treatment of firearms examination. *Id.*

***