UCLA
UCLA Criminal Justice Law Review

Title
The (Im)Proficiency of Firearms Examiners

Permalink
https://escholarship.org/uc/item/0gk6r79s

Journal
UCLA Criminal Justice Law Review, 10(1)

Author
Gutierrez, Richard E.

Publication Date
2026-11-16

DOI
10.5070/CJ8.70319

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

THE (IM)PROFICIENCY OF FIREARMS EXAMINERS

Richard E. Gutierrez

ABSTRACT

For decades, the nation's largest provider of forensic proficiency tests, Collaborative Testing Services, structured its firearms examination exercises around bullet and cartridge case comparisons so simplistic as to nearly guarantee vanishingly low misidentification rates. Practitioners, in turn, enthusiastically relied on these results during increasingly vigorous admissibility challenges to defend their alleged ability—using little more than their training, experience, and subjective judgment—to divine the precise gun used in a crime. By and large, they succeeded. Although outside scholars (lawyers, statisticians, human factors experts, and more) have remained constant in their criticisms of the field's methods and the design of its efforts to measure accuracy (in the form of proficiency testing and beyond), untroubled admissibility and untroubled qualification of firearms examiners as "experts" have remained the norm. In 2023, however, CTS (albeit unintentionally) bucked the trend of simplistic testing, releasing an exercise, CTS 23–5262, that forced participants to confront damaged and poorly marked bullets. The results should trouble even those unfamiliar with firearms examination and the debates about its validity. Although the questioned bullets provided to participants were not fired by the same gun as the exemplar bullets in their packets, 22% of test takers failed by "matching" at least one questioned item to the exemplars, and participants on the whole committed misidentifications 37.4% of the time they offered definitive conclusions about the source of the "challenging" bullets in the proficiency test. To highlight the importance of these results, and to emphasize the extent they depart from past proficiency tests, this Article both analyzes CTS bullet exercises dating back to 2006 and dives deeply into CTS 23–5262 itself. Examiners' performance failures on the latter, this Article ultimately argues, (1) contradict claims that analysts will rarely err when assessing the source of fired ammunition; (2) raise questions about the validity of comparisons performed on damaged bullets; and (3) force litigants and the courts to grapple with the significance of test error when qualifying individual examiners as experts. Indeed, given how grossly CTS 23–5262 runs counter to the claims made by defenders of firearms examination (and the judges who have parroted them), it may be that the test can (or at least should) finally push the field over the edge and out of criminal courts.

© 2026 Richard E. Gutierrez. All rights reserved.

About the Author

Richard E. Gutierrez is an Assistant Professor of Law at the University of Illinois Chicago (UIC) School of Law. He received his B.A. from the University of Florida and his J.D. from Chicago-Kent College of Law. He is grateful to Emily Prokesch and Joseph Cavise for the many discussions they endured about this paper and the excellence of their feedback. He also thanks the Chicagoland Junior Scholars Conference and the Central States Scholarship Conference for the opportunity to present drafts of this work (both early and late in its development). Finally, he extends heartfelt acknowledgment to Emma Armbrust, who provided excellent research assistance and made his first experience working with a student law aid an easy and fruitful one.

Table of Contents

Introduction 93

I. Empiricism Versus Tradition: The State of the Debate Surrounding Firearms Examination & Proficiency Testing 101

A. *The Uncharted Wilds of Firearms Examination* 105

B. *Proficiency Tests & Consumer Satisfaction, A Recipe for Meaningless Results* 114

II. Singularly Formidable & Fluky: CTS 23–5262's Predecessors, Samples, Participants, & Performance Measures 121

III. Having Heard the Bell Rung: The Path Forward for Litigants & the Courts 133

A. *Reassessing Assumptions About False Positive Rates* 135

B. *Validity As Applied for Damaged Bullets* 142

C. *Separating the Wheat from the Chaff & Exploring Qualifications At & Before Trial* 148

Conclusion 157

Introduction

Back in 2017, while still working as a member of the Law Office of the Cook County Public Defender's specialized Forensic Science Division,[1] I got the chance to attend a symposium hosted by the Bluhm Legal Clinic's Center on Wrongful Convictions covering "Hot Topics in Criminal Justice."[2] And while there was a lot on offer for a young-ish public defender to get excited about (even my boss got a turn on a panel discussing implicit bias), what really sparked my interest was a talk by a longtime proponent and practitioner of firearms examination, John Murdock.[3] For years, blue-ribbon government panels had repeatedly questioned the validity of the bullet and cartridge case comparisons performed by firearms examiners.[4] And I had, in turn, spent the bulk of my time filing motions challenging Illinois's longstanding acceptance[5] of such evidence. At the symposium, I had the opportunity to engage directly with a man who had spent a career on the other side, defending the accuracy and admissibility of a discipline[6] somehow wholly premised on the subjective comparison of microscopic scratches and dents, and yet (allegedly) still capable of identifying the specific gun responsible for discharging relevant ammunition

1. *Forensic Science Division*, Law Off. Cook Cnty. Pub, Def., https://www.cookcountypublicdefender.org/forensic-science-division [https://perma.cc/TX84-LZJ8] (last visited Apr. 7, 2026).
2. Nw. Pritzker Sch. of L. Bluhm Legal Clinic, *Center on Wrongful Convictions*, News & Notes, Fall 2017, at 17–18. No recordings of the symposium nor powerpoint presentations are available online. I have therefore reconstructed, as faithfully as possible, my account from memory.
3. *Id.* This article will return to describe the methods and criticisms of firearms examination, *see infra* Section I.A, but for now the idea here is that marks "produced by different firearms vary substantially enough (owing to variations in manufacture and use) to allow components of fired cartridges to be identified with particular firearms," by trained firearms examiners using their subjective judgement. President's Council of Advisors on Sci. & Tech., Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods 11 (2016) [hereinafter PCAST Report].
4. PCAST Report, *supra* note 3, at 11 (Firearms examination "falls short of the scientific criteria for foundational validity."); Comm. On Identifying the Needs of the Forensic Sci. Cmty., Nat'l Rsch. Council, Strengthening Forensic Science in the United States: A Path Forward 7 (2009) ("[N]o forensic method [besides DNA analysis] 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.") [hereinafter NAS Report]; Nat'l Rsch. Council, Ballistic Imaging 82 (Daniel L. Cork et al. eds., 2008) ("[T]he validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated.") [hereinafter NRC Report].
5. *See* People v. Fisher, 172 N.E. 743 (Ill. 1930); People v. Robinson, 2 N.E.3d 383 (Ill. App. 1st Dist. 2013); People v. Rodriguez, 106 N.E.3d 436 (Ill. App. 1st Dist. 2018).
6. *See, e.g.*, Richard Grzybowski et al., *Firearm/Toolmark Identification: Passing the Reliability Test Under Federal and State Evidentiary Standards*, 35 AFTE J 209 (2003); Richard A. Grzybowski & John E. Murdock, *Firearm and Toolmark Identification - Meeting the* Daubert *Challenge*, 30 AFTE J 3 (1998); Alfred A. Biasotti & John Murdock, *"Criteria for Identification" or "State of the Art" of Firearm and Toolmark Identification*, 16 AFTE J 17 (1984).

with certainty.[7] But I wasn't thrilled at the chance to pick a fight. I was excited because I thought that Mr. Murdock had likely received his invitation to speak not based on that scholarship, but instead given his role in uncovering one of those errors his field so confidently dismissed as a "practical impossibility:"[8] the misidentification of bullets and cartridge cases by the Illinois State Police Division of Forensic Services that had led directly to the wrongful conviction of Patrick Pursley for first-degree murder.[9] I thought I might get to witness an about-face, that uncovering a miscarriage of justice which had cost an innocent man twenty-four years of freedom might have swayed Mr. Murdock in a way that years of weathering criticism of his chosen field by outside scholars[10] had not.

After all, Mr. Pursley's story was more than just tragic. It was, in my mind, an indictment that cut to the core of firearms examination.[11] Here was an error that could neither be dismissed as the product of some renegade or underperforming laboratory,[12] nor laid at the feet of defense attorneys unable or unwilling to mount an attack on complex scientific evidence.[13] Instead,

7. *See* Comm. for the Advancement of the Sci. of Firearm & Toolmark Identification, *Theory of Identification as It Relates to Toolmarks: Revised*, 43 AFTE J. 287 (2011).
8. *Id.*
9. *Patrick Pursley*, Nw. Pritzker Sch. L. Bluhm Legal Clinic Ctr. on Wrongful Convictions, www.law.northwestern.edu/legalclinic/wrongfulconvictions/exonerations/patrick-pursley.html [https://perma.cc/D75S-W3MR] (last visited May 22, 2026).
10. *See infra* Part I.A.
11. *See* John Morgan, *Wrongful Convictions and Claims of False or Misleading Forensic Evidence*, 68 J. Forensic Sci. 908, 949 (2023) (dismissing many of the wrongful convictions associated with firearms examination evidence as resulting from dated approaches or poor laboratory management but saying "[t]he Pursley case is unlike any other firearms identification case in the dataset. All other misidentifications can be associated with training and practice deficiencies. The six Pursley examiners largely conformed to consensus standards, although Gunnell overstated the statistical weight of his conclusion at trial").
12. Indeed, the Illinois State Police lab was the first in the world to achieve accreditation back in 1982. *See* Ill. State Police, Annual Report 14 (2012); *see* Nat'l Comm'n on Forensic Sci., Recommendation to the Attorney General: Universal Accreditation 6 (2015) ("Accreditation is an independent third-party assessment of a [laboratory]'s . . . quality, administrative, and technical systems. Accreditation uses specific criteria and procedures based upon accepted standards to ensure the quality of the [laboratory]'s management system").
13. Not only did Mr. Pursley's attorneys hire, however unfruitfully, a firearms examiner to review the work of the Illinois State Police, they also undertook a thorough cross examination of the state's expert witness. *See* People v. Pursley, 943 N.E.2d 98, 107–08 (Ill. App. Ct. 2011). That Pursley's conviction occurred despite such able representation should terrify defendants whose liberty forensic evidence imperils. Law students and practicing attorneys face a dearth of educational opportunities to prepare them for cases involving forensic evidence. *See* Brandon L. Garrett et al., *Forensic Science in Legal Education*, 51 J. Law & Educ. 1, 12 (2022); Robert M. Sanger, *Forensics: Educating the Lawyers*, 43 J. Legal Pro. 221, 250 (2019); 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 Sci. 1334, 1345–46 (2024). And scholars have tied that lack of training (as well as deficient resources) directly to less-than-impeccable litigation. *See e.g.*, Mark Loudon-Brown, *Garbage In, Garbage*

multiple firearms examiners had simply gotten it wrong.[14] They had transformed an incredibly weak case into a conviction by falsely linking a gun from Mr. Pursley's home to bullets and cartridges from the murder scene,[15] thereby caging an innocent man who very nearly (given his sentence and failed appeals) never again tasted liberty.[16] So yes, Mr. Murdock had bristled to see attorneys lambasting his discipline.[17] Sure, he had explained away those blue-ribbon

Out: Revising Strickland As Applied To Forensic Science Evidence, 34 Ga. St. U. L. Rev. 893, 895(2018); Peter J. Neufeld, *The (Near) Irrelevance of* Daubert *to Criminal Justice and Some Suggestions for Reform*, 95 Am J. Pub. Health 107, 110 (2005).

14. *See* People v. Pursley, No. 2–17–0227, 2018 WL 2095855, at *6, *8–10 (Ill. App. 2 Dist. May 2, 2018)(At least three Illinois State Police examiners and the defense's own expert failed to exclude Pursley's gun as the source of the crime scene bullets and cartridge cases despite Mr. Murdock going so far as to say that he did not even take photographs because "there was really nothing to illustrate. There was no significant agreement at all.").

15. The surviving eyewitness could not identify Mr. Pursley; the individuals the State called to prove that he had admitted guilt fared no better with one admitting to receiving $2,650 to incriminate Pursley, and the other recanting at trial claiming that police had coerced her into providing an inculpatory statement; and his defense team managed to present substantial exculpatory evidence including three alibi witnesses, as well as an eyewitness who described the killer as a full five inches taller than Mr. Pursley. *Id.* at *2; *see also* Maurice Possley, *Patrick Pursley*, Nat'l Registry Exonerations (Apr. 29, 2023), https://exonerationregistry.org/cases/12582. That firearms examination evidence could exert such influence is not surprising. As discussed further below, *see infra* Part III.A., lay people grossly underestimate the potential for errors by firearms examiners, are willing to convict based on little (if nothing) other than the testimony of a firearms examiners, and do not much moderate their views based even on effective cross examination. *See e.g.*, Brandon L. Garrett et al., *The Impact of Defense Experts on Juror Perceptions of Firearms Examination Testimony*, 64 Jurimetrics 223, 241–42 (2024) (finding that "a sizable majority of [lay people] (72%) believed the false positive rate [for firearms examination] to be orders of magnitude smaller than" error rate study estimates, and that most "(66%) were also willing to convict on the basis of firearms examination evidence alone when it went unrebutted"); Brandon L. Garrett et al., *Mock Jurors' Evaluation of Firearm Examiner Testimony*, 44 Law & Hum. Behav. 412, 421–22 (2020) (finding that cross examination largely did not impact guilty verdicts based on firearms examination evidence).

16. *Pursley*, 2018 WL 2095855, at *2–4 (Pursley originally received a sentence of life in prison without the possibility of parole, or only obtained release and a new trial to prove his innocence after multiple postconviction petitions). Pursley eventually obtained an acquittal and a certificate of innocence. *See* Ken DeCoster, *Patrick Pursley Officially Declared Innocent of 1993 Murder in Rockford*, Rockford Reg. Star (Feb. 26, 2021), https://www.rrstar.com/story/news/2021/02/26/patrick-pursley-gets-certificate-innocence-after-wrongful-conviction/6814469002/ [https://perma.cc/YU9P-5YTQ]. Though other consequences pale in comparison to what Mr. Pursley suffered, the taxpayers of Illinois also paid quite dearly for his wrongful conviction. *See* Jeff Kolkey, *Rockford Settles Wrongful Conviction Lawsuit for $5M*, Rockford Reg. Star (Dec. 19, 2024), https://www.rrstar.com/story/news/crime/2024/12/19/rockford-to-pay-patrick-pursley-5m-to-settle-wrongful-conviction-suit/77074374007/?gnt-cfr=1&gca-cat=p&gca-uir=true&gca-epti=z118707e002400v118707d—87—b—87—&gca-ft=143&gca-ds=sophi [https://perma.cc/ZC9K-UGW9].

17. *See, e.g.*, Adina Schwartz, *A Systemic Challenge to the Reliability and Admissibility of Firearms and Toolmark Identification*, 6 Colum. Sci. & Tech. L. Rev. 2 (2005). Although, to my knowledge, Mr. Murdock did not reply to Professor Schwartz in writing to reject her arguments, his peers did. *See, e.g.*, Ronald G. Nichols, *Defending*

government reports[18] as merely the product of a lack of familiarity with discipline literature.[19] But finding an error himself, at the hands of his field's best no less—that, I naively thought, had to change things.

I could not have been more wrong. Over the course of an hour, Mr. Murdock launched a full-throated defense of firearms examination, refused to grapple with the consequences of an error that took place just down the road from the podium he stood behind, and never so much as mentioned Mr. Pursley by name. What really struck me, however, was how he ended that talk. With an audience eager and antsy to push back (at least I was), Mr. Murdock blew right through the time for questions. And when he advanced to his final slide, all he showed that audience was an image of a gun, pointing directly at us and going off. What he said shocked me even more: "If this is the last thing you see before you die, don't worry: we'll be able to get the guy!" That image, and those words, have haunted me across the many years since. Their overflowing self-assurance came back to mock me when I finally succeeded in persuading a judge to bar firearms examination evidence,[20] only to see that fragile victory ripped away a few months later by his successor.[21] They jumped to mind immediately when I read another firearms examiner's report uncovering additional misidentifications from an accredited lab (this time in Rhode Island) while

---

the Scientific Foundations of the Firearms and Tool Mark Identification Discipline: Responding to Recent Challenges, 52 J. Forensic Sci. 586 (2007).

18. *Supra* note 3 & *supra* note 4.

19. *See, e.g.*, John Murdock et al., *The Response of the Association of Firearm and Tool Mark Examiners to the February 2009 National Academy of Science Report "Strengthening Forensic Science in the United States: A Path Forward"*, 41 AFTE J 204 (2009) ("the NAS Committee in effect chose to ignore extensive research supporting the scientific underpinnings of the identification of firearm and toolmark evidence"); John E. Murdock *et al*., *The Development and Application of Random Match Probabilities to Firearm and Toolmark Identification*, 62 J. Forensic Sci. 619, 619 (2017) ("The field of firearms and toolmark analysis has encountered deep scrutiny of late, stemming from a handful of voices, primarily in the law and statistical communities. While strong scrutiny is a healthy and necessary part of any scientific endeavor, much of the current criticism leveled at firearm and toolmark analysis is, at best, misinformed and, at worst, punditry.").

20. Illinois v. Winfield, No. 15CR14066–01 (Cir. Ct. Cook Cnty. Feb. 8, 2023) (on file with author).

21. Illinois v. Winfield, No. 15CR14066–01 (Cir. Ct. Cook Cnty. Dec. 18, 2024) (on file with author). Even this ruling was never truly finalized given that the defense filed its own motion to reconsider that reversal that remained pending at the time the defendant plead guilty. *See* Motion to Reconsider the People of Illinois' Motion to Reconsider the Exclusion of Firearms Testimony, *Illinois v. Winfield*, No. 15CR14066–01 (Cir. Ct. Cook Cnty. Jan. 16, 2025) (on file with author). For those interested in this case, or about the fragility of victories against faulty forensic sciences more generally, Radley Balko has done excellent reporting. *See* Radley Balko, *Devil in the Grooves: The Case Against Forensic Firearms Analysis*, The Watch (May 25, 2023), https://radleybalko.substack.com/p/devil-in-the-grooves-the-case-against [https://perma.cc/DLN4–4WPC]; Radley Balko, *A Chicago Judge Just Erased Her Predecessor's Historic Ruling on Forensic Firearms Analysis*, The Watch (Jan. 14, 2025), https://radleybalko.substack.com/p/a-chicago-judge-just-erased-her-predecessors [https://perma.cc/7KJJ-ZD9Q]..

simultaneously laying out an unequivocal defense of the discipline.[22] And they were the very first thing that I thought of when I came across CTS 23–5262, a bullet comparison exercise that, albeit unintentionally, (1) included challenging samples, poorly marked bullets often damaged by "tank rash"; (2) duped more than 20% of participating examiners into misidentifications; (3) kept all but a third of participating examiners from diagnosing its more challenging comparisons as involving bullets fired by different guns; and (4) resulted in a false positive (misidentification) rate almost no better than flipping a coin would produce.[23]

It was only my discovery of CTS 23–5262, however, that brought on that same curiosity that had driven me to hear Mr. Murdock speak back in 2017. Would the test's results, given their gross inconsistency with examiner claims of certainty and law enforcement's vanishingly low estimates of the potential for error,[24] finally be the thing to shake him, to pull him back from the kind of swaggering confidence necessary to point a gun (however metaphorical) at an audience full of defense and innocence attorneys and promise them post-mortem revenge? It wasn't a foolish thought. Though they had long been criticized as meaninglessly easy and simplistic,[25] Mr. Murdock had repeatedly relied[26] on other Collaborative Testing Services (CTS) proficiency exams to establish error rates satisfying Federal Rule of Evidence 702 and similar state equivalents.[27] But given the rush by other firearms examiners to dismiss CTS

22. *See* Ronald Nichols, Report of Findings Rhode Island State Crime Laboratory Firearm and Toolmark Examination Unit 1–3, 6 (Oct. 9, 2024) (spending 2.5 pages defending the "Validity of Forensic Firearm and Toolmark Examination" before conceding that one of the errors committed by multiple examiners occurred despite obvious differences; *see also* Mark Reynolds, *Expert Says RI Crime Lab in Need of Reform After Troubling Findings on Bullet Examination*, Providence J. (Oct. 24, 2024), https://www.providencejournal.com/story/news/crime/2024/10/24/forensic-experts-report-exposes-failures-at-ri-crime-lab/75814516007/; Michael A. DiLauro, *Lessons Learned From A Not So Public 'Mismatch': Firearms, Toolmark Analysis, and Rhode Island's Publicly Funded Crime Laboratories*, R.I. Bar J. 13 (2025).
23. *See* Collaborative Testing Servs., Inc., Firearms Examination Test No. 23–5262 Summary Report (2024) [hereinafter CTS 23–5262]. For a breakdown of these figures and additional detail on CTS 23–5262, see *infra* Part II.
24. *See supra* note 6; Jim Agar, *The Admissibility of Firearms and ToolMarks Expert Testimony in the Shadow of PCAST*, 74 Baylor L. Rev. 93, 193 (2022) (claiming a false positive rate for firearms examination under 2%); FBI, *Response to the Declaration Regarding Firearms and Toolmark Error Rates Filed in* Illinois v. Winfield, May 3, 2022, at 3 (on file with author) ("studies produced low false positive rates of approximately 1% or less"); Richard E. Gutierrez, *Bowling with Bumper Rails: How Firearms Examiners Have Duped the Courts and Generated Low Error Rates Only by Avoiding Challenging Comparisons*, 75 U.C. L.J. 1535, 1549 & 1555 (2024) (documenting judicial acceptance of such figures).
25. *See infra* Part I.B.
26. *See* Grzybowski et al., *supra* note 6, at 216–19; Grzybowski & Murdock, *supra* note 6, at 9.
27. *See* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 593–95 (1993); *see also* 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) (reviewing varying admissibility standards for scientific evidence).

23–5262's results as inconsequential,[28] I doubt the test would have reined in his claims. Maybe, as courts have long cautioned, there is just no convincing insiders whose "reputation and livelihood depen[d] on use of [a] technique," and just no reason to expect that practitioners would do anything other than "in effect self-certify the validity" of their own methods.[29]

Yet whatever its ability to change the hearts and minds of firearms examiners, CTS 23–5262 is ideally positioned to echo robustly in the courts. For years, the field of firearms examination has failed to get itself off the ropes. Scholars have largely discounted existing estimates of error in the discipline as meaningless.[30] Training and experience appear, at best, to have a troubled relationship to superior performance.[31] Multiple studies have produced double-digit rates of misidentification.[32] And studies that show rosier likelihoods of

28. *See infra* Part II; James Carroll et al., *Report of the Association of Firearm and Toolmark Examiners Proficiency Test Review Ad-Hoc Committee on the Results of Collaborative Testing Service (CTS) Firearms Examination Proficiency Test 23–5262*, 56 AFTE J 129 (2024).

29. People v. Young, 391 N.W.2d 270, 276 n.24 (Mich. 1986); *see also Ramirez v. State*, 810 So.2d 836, 844 n.13 (Fla. 2001) (warning against reliance on experts who have a "personal stake" in the acceptance of a methodology or show "institutional bias"); People v. Wesley, 633 N.E.2d 451, 464 (N.Y. Ct. App.1994) (Kaye, J., concurring) ("A *Frye* court should be particularly cautious when—as here—"the supporting research is conducted by someone with a professional or commercial interest in the technique" (citation omitted)); State v. Williams, 147 N.E.3d 1131, 1142 (N.Y. Ct. App. 2020) (admission must be "supported by those with no professional interest in its acceptance. *Frye* demands an objective, unbiased review." (footnote omitted)); Almeciga v. Ctr. For Investigative Reporting, Inc., 185 F. Supp. 3d 401, 420 (S.D.N.Y. 2016) (noting that "'review by a "competitive, unbiased community of practitioners and academics"' . . . would be expected in the case of a scientific field" (citations omitted); United States v. Porter, 618 A.2d 629, 634 (D.D.C. 1992) ("It simply is not creditable to argue . . . that general acceptance may be premised simply on the opinion of forensic scientists." (citation omitted)).

30. *See e.g.*, Maria Cuellar et al., *Methodological Problems in Every Black-box Study of Forensic Firearm Comparisons*, 23 L. Prob. & Risk 1, 1 (2024) ("]E]rror rates among firearms examiners, both collectively and individually, remain unknown. Therefore, statements about the common origin of bullets or cartridge cases that are based on examination of "individual" characteristics do not have a scientific basis."); *see infra* Part I.A.

31. *See, e.g.*, Richard E. Gutierrez, *More Unjustified Inferences From Limited Data in Guyll et al. (2023)*, 23 L. Prob. & Risk 1, 1 (2024) (noting that one firearms study failed to "correlate[e] performance in some way to years of experience, type of training received, laboratory accreditation, certification, or the like"); Richard E. Gutierrez & Celeste Addyman, *Commentary on: Monson KL, Smith ED, Peters EM. Accuracy of Comparison Decisions by Forensic Firearms Examiners*, 68 J. Forensic Sci. 1097, 1100 (2023) ("[J]udges . . . have precious little (if any) means . . . to assess . . . whether a given examiner is qualified to offer expert testimony: none of training, experience, or accreditation appear to predict accuracy . . . . ").

32. *See* Gutierrez, *supra* note 24, at 1539–40 (*citing* Julie Knapp & Angela Garvin, *Consecutively Manufactured .25 Auto F.I.E. Barrels- A Validation Study*, Presentation at AFTE 43rd Annual Training Seminar (2012)); Petra Pauw-Vugts et al., *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.

error have been revealed to utilize almost laughably simplistic samples.[33] Indeed, after facing two decades of sustained criticism from outside scholars,[34] the discipline remains so short on disinterested allies that one judge quipped that "rarely do the experts fall into such cognizable camps, forensic practitioners on one side and academic researchers on the other."[35] Despite all this, admissibility of evidence from the field remains the norm; but increasingly robust challenges from defense attorneys have recently sparked "a dramatic rise in a judge's willingness to engage with scientific limitations of [firearms examination] methods."[36] And more than just engage, multiple trial judges from across the United States have, in the last few years, strictly limited, if not excluded outright, testimony of firearms examination "experts,"[37] as have two state appellate courts.[38] Still, though a plethora of convincing evidence supported knocking firearms examination out of its privileged position in the criminal legal system even before CTS 23–5262, the test really does hit examiners where it hurts most and strikes at the heart of the defenses that have won them favor among judges. It validates longstanding concerns about the ease of proficiency tests[39] (as well as error rate studies more generally)[40] and troubles the judiciary's repeated reliance on them to admit bullet and cartridge case comparison testimony;[41] contradicts discipline claims and judicial findings about the potential misidentification rate associated with firearms examination;[42] and calls deeply into question the ways that firearms examiners have survived qualification inquiries through little more than declarations of "training and

Mattijssen et al., *Validity and Reliability of Forensic Firearm Examiners*, 307 Forensic Sci. Int'l 110, 112 (2020).

33. *See generally* Gutierrez, *supra* note 24, at 1561–73 (outlining the simplicity of comparison samples and their failure to mimic the difficulties of casework across all major error rate studies); *see also* Gutierrez & Prokesch, *supra* note 13, at 1334 (one study used comparison samples so obviously dissimilar that lay people were able to perform just as well as professional examiners on different-source samples).
34. *See infra* Part I.A.
35. People v. Ross, 129 N.Y.S.3d 629, 639 (N.Y. Sup. Ct. 2020).
36. Brandon L. Garrett et al., *Judging Firearms Evidence*, 97 S. Cal. L. Rev. 101, 109 (2023) (citation omitted).
37. Gutierrez, *supra* note 24, at 1538 (collecting cases). Indeed, even since that article's list, another trial judge in Chicago has excluded outright source attribution testimony from a firearms examiner. *See* People v. Kimberley, No. 22-CR-02932–01 (Cir. Ct. Cook Cnty. Jul. 1, 2025) (on file with author).
38. State v. Adams, 572 P.3d 291, 313 (Or. App. Ct. 2025) ("We therefore conclude on this record that the state did not carry its burden of showing that the AFTE method is scientifically valid."); Abruquah v. State, 296 A.3d 961, 997 (Md. 2023) ("[W]e conclude that the methodology of firearms identification presented . . . did not provide a reliable basis for . . . [an] unqualified opinion that four bullets and one bullet fragment found at the crime scene in this case were fired from`[the defendant's] Taurus revolver.").
39. *See infra* Part I.B.
40. *See e.g.*, Gutierrez, *supra* note 24, at 1562 ("[F]irearms examination has largely refused to explore performance on . . . challenging samples.").
41. *See infra* Part III.A.
42. *See id.*

experience."[43] Given the results of CTS 23–5262, firearms examination's past cannot be its prologue. Judges should not allow the field to retreat to near-perfect performance born only of simplistic testing,[44] or to survive and thrive despite its inability to silence its myriad doubters.[45] Instead, if the veneer of firearms examination's infallibility cannot withstand run-ins with comparisons actually resembling the real world, as CTS 23–5262 shows, then the discipline's century-long entanglement with our criminal legal system must end. It is time, at last, to nail the coffin shut, if not on firearms examination writ large, then at least on the comparison of damaged bullets.

To ensure that CTS 23–5262 has that robust impact, that its chilling consequences do not suffer the same dustbin of history fate as so much of the other evidence arrayed against firearms examination, this Article takes up the task of documenting the details of that exam, laying bare the implications of practitioners' widespread failures to rise to meet its challenges, and urging courts to break with the longstanding and unjustified tradition of admitting firearms examination evidence and qualifying its practitioners without searching review. Part I provides background on the results from CTS 23–5262, detailing the basics of firearms examination and the depth of criticisms leveled at its validity, as well as laying out the promise, present reality, and underperformance of existing proficiency testing regimes. Part II deconstructs CTS 23–5262 itself, first comparing the test to its predecessors to emphasize the extent of its departure from past practice, and next breaking down the creation of the test, the characteristics of its samples, and the disturbing failures of its participants. And finally, Part III addresses the ways that CTS 23–5262 should impact litigants and the courts, urging them to reconsider beliefs about firearms examination's potential to provoke wrongful convictions, question the application of firearms examination methods to bullet comparisons, and grapple fully with the qualifications of examiners who failed to navigate the test's comparisons.

43. *See* Garrett et al., *supra* note 36, at 153 ("Many courts traditionally focused on an expert's credentials and self-professed expertise when conducting this inquiry into the qualifications of the witness." (citation omitted)); Brandon L. Garrett & Gregory Mitchell, *The Proficiency of Experts*, 166 Univ. Penn. L. Rev. 901, 903 (2018) ("[C]ourts routinely accept a witness's own self-serving statements of expertise buoyed by educational credentials, professional training, or experience, rarely spending much time on this threshold question [of a witness's qualifications] before moving on to examine the methods used and conclusions reached by the putative expert." (citation omitted)).

44. Indeed, CTS has already signaled its intention, going forward, to "meet industry needs and customer expectations" (i.e. to shy away from tests like CTS 23–5262 and to instead produce those easier tests forensic labs and examiners prefer). CTS 23–5262, *supra* note 23, at "frequently asked questions" 5.

45. *See* Garrett & Mitchell, *supra* note 43, at 93 (arguing that so long as judges continue to accept minimal testing of forensic disciplines and practitioners, or flip the burden onto opponents of such evidence, "forensic evidence professions have little incentive to engage in rigorous proficiency testing").

I. Empiricism Versus Tradition: The State of the Debate Surrounding Firearms Examination & Proficiency Testing

To understand why CTS 23–5262 should so impact judicial treatment of firearms examination, readers uninitiated into forensic science's history in the courts need something of a crash course in the basics of what expertise even means. Most Americans are confident in scientific experts, in the need for scientific research, and in the belief that scientists will "act in the best interests of the public."[46] But if grounded in blind deference to supposed expertise, rather than data on superior performance, our reliance on specialists (scientific or otherwise)[47] can quickly lead us astray.[48] If you have ever felt a sense of shame because your taste in wine diverges from what sommeliers recommend, it might be time to take it easy on yourself: "expert" wine tasters, among other failings, rarely provide consistent ratings, describe the same wine differently depending on what its label suggests about quality and price, and have even been duped by food dye into describing white wines as reds.[49] If you are depending on a wealth manager to help you strike it rich, you might be better served seeking help from your pets instead.[50] And if you are paying a premium

46. *See, e.g.*, Nat'l Sci. Bd., Science and Technology: Public Perceptions, Awareness, and Information Sources 5 (2024); Carissa Wong, *Largest Post-Pandemic Survey Finds Trust in Scientists is High*, Nature (Feb. 14, 2024), https://www.nature.com/articles/d41586–024–00420–1 [https://perma.cc/4M8N-X48L]; Jon D. Miller et al., *Citizen Attitudes Toward Science and Technology, 1957–2020: Measurement, Stability, and the Trump Challenge*, 51 Sci. & Pub. Policy 526 (2024); *but see* Brian Kennedy & Alec Tyson, *Americans' Trust in Scientists, Positive Views of Science Continue to Decline*, Pew Rsch. Ctr., 3–7 (2023) (concluding that, although most Americans continue to view science and scientific experts positively, the percentage of Americans holding that view has declined since 2019, especially among Republicans).

47. Confidence in experts more generally may be lower than in scientists specifically. *See* Kennedy & Tyson, *supra* note 46, at 4 ("[S]cientists and medical scientists continue to be held in high regard compared with other prominent groups in society. Smaller shares of Americans express confidence in business leaders, religious leaders, journalists and elected officials to act in the public's best interests. As with scientists, most of these groups have seen their ratings decline in recent years."); Andrew J. Taylor, *The Expert in American Life*, 60 Nat'l Affairs (2021) (noting survey in which 60% of Republicans and 44% of Democrats agreed that they would "rather put [their] trust in the wisdom of ordinary people than the opinions of experts and intellectuals"), https://www.nationalaffairs.com/publications/detail/the-expert-in-american-life [https://perma.cc/6KRR-25PZ].

48. *See e.g.*, K. Anders Ericsson et al., *The Making of an Expert*, Harv. Bus. Rev. (2007) (noting many "fields where there is no scientific evidence that supposed expertise leads to superior performance"), https://hbr.org/2007/07/the-making-of-an-expert [https://perma.cc/M248-QZBV].

49. *See id.*; David Derbyshire, *Wine-Tasting: It's Junk Science*, The Guardian (Jun. 22, 2013), https://www.theguardian.com/lifeandstyle/2013/jun/23/wine-tasting-junk-science-analysis [https://perma.cc/6TWF-G6ML]; Ross Pomeroy, *The Legendary Study that Embarrassed Wine Experts Across the Globe*, Real Clear Sci. (Aug. 19, 2014), www.realclearscience.com/blog/2014/08/the_most_infamous_study_on_wine_tasting.html [https://perma.cc/5LXK-TQ33].

50. *See* Mandi Woodruff, *This Cat Can Pick Stocks Better Than Most Investment Managers*, Yahoo Fin. (Jan. 14, 2013), https://finance.yahoo.com/news/cat-pick-stocks-better-most-192600625.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAEgeL-

for your chosen therapist or physician because of their "advanced degrees and decades of experience," you could just as well save some money by hiring a replacement fresh out of school (you might even get better results).[51] For these reasons, social scientists and legal scholars have defined expertise not by reference to surrogate measures (like training, acclaim from peers, knowledge of a domain, years of experience, or other credentials), but instead empirically and as a function of superior performance: demonstrable success beyond that of novices on domain tasks.[52]

What is true for science and expertise more generally applies with equal force to forensic[53] practitioners and their methodologies. Courts, in line with research on lay reactions to forensic and other expert testimony,[54] have recognized that "[e]xpert evidence can be both powerful and quite misleading,"[55] have highlighted the "natural inclination of the jury to equate science with truth and, therefore, *accord* undue significance to any evidence labeled scien-

rfHYp49y0Duhkwajc07ua6nlZyCKqlziXJfiw_C3HBKx9aCSmuUaychF_KXZhK-W7CwmXkIc6MrFXEJMCYkkvWqLapVV3gvKfw91YVlO54FOTQNtIA0D-NVsOTyckQ1PWGflLjffoAOKZewCURAgw_Y5vOJuFUP2ESCoM0tsr; K. Anders Ericsson & Neil Charness, *Expert Performance: Its Structure & Acquisition*, 49 Am Psych. 725, 731 (1994) ("[S]tudies have found that financial experts' stock investments yield returns that are not consistently better than the average of the stock market, that is, financial experts' performance does not differ from the result of essentially random selection of stocks.").

51. *See* Ericsson et al., *supra* note 48 (noting that "psychotherapists with advanced degrees and decades of experience aren't reliably more successful in their treatment of randomly assigned patients than novice therapists with just three months of training are" and that "[t]he longer physicians have been out of training . . . the less able they are to identify unusual diseases of the lungs or heart").

52. *See e.g.*, James Shanteau et al., *How Can You Tell If Someone is an Expert? Performance-Based Assessment of Expertise*, Emerging Persps. on Judgment & Decision Rsch. 620, 622–27 (Sandra L. Schneider & James Shanteau, eds., 2003) (criticizing other approaches to evaluating experts including experience, training, and referring to them as, at best, "uncertain predictor[s] of degree of expertise"); David J. Weiss & James Shanteau, *Decloaking the Privileged Expert*, 18 J. Mgmt. & Org. 300, 308 ("We view education and training as precursors rather than as guarantors of expertise."); David J. Weiss & James Shanteau, *Empirical Assessment of Expertise*, 45 Hum. Factors 104, 104 (2003) ("Experts have often been identified by self-proclamation or acclamation by other experts as well as by experience, titles, and degrees. However, these methods can be misleading when searching for an expert. We prefer instead to cast the problem in empirical terms: An expert is someone who carries out a specified set of tasks expertly."); Ericsson & Charness, *supra* note 50, at 731 ("The distinction between the perception of expertise and actual expert performance becomes increasingly important as research has shown that the performance of some individuals who are nominated as experts is not measurably superior."); Garrett & Mitchell, *supra* note 43, at 904, 910 (opining that "anyone who claims to be an expert should be able to prove that expertise through superior performance").

53. Nat'l Inst. Standards & Tech., Strategic opportunities to Advance Forensic science in the United States: A Path Forward Through Research & Standards iv (2024) ("Forensic science is the application of scientific or technical practices to the recognition, collection, analysis, and interpretation of evidence for criminal and civil law or regulatory issues." (citation omitted)).

54. *See infra* Section III.C.

55. *Daubert*, 509 U.S. at 595 (citation omitted).

tific,"[56] and have even described experts as wielding "talismanic significance" in the eyes of jurors.[57] As the President's Council of Advisors on Science and Technology (PCAST) to Obama emphasized:

> [w]ithout appropriate estimates of accuracy, an examiner's statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact. Nothing—not training, personal experience nor professional practices—can substitute for adequate empirical demonstration of accuracy.[58]

Forensic practitioners, however, chafe against calls to quantify accuracy, manipulate the language of scientific theories, lean heavily on "training and experience," and cite the difficulty of producing useful error rates.[59] But when subjected to empirical scrutiny, forensic professionals (from bitemark examination to DNA) also fail to live up to the hype with alarming frequency and

56. In re Det. of New, 21 N.E.3d 406, 412 (Ill. 2014) (citation omitted).

57. United States v. Frazier, 387 F.3d 1244, 1263 (11th Cir. 2004). The courts just cited are far from alone in expressing concerns regarding the extent jurors will defer to expert witnesses. *See infra* Section III.A. And scientific panels have gone even further when discussing typical forensic experts. *See e.g.*, PCAST Report, *supra* note 3, at 45 (concluding that that "[c]ompared to many types of expert testimony, testimony based on forensic feature-comparison methods poses unique dangers of misleading jurors," because "[t]he vast majority of jurors have no independent ability to interpret the probative value of results based on the detection, comparison, and frequency of scientific evidence . . . they would be completely dependent on expert statements garbed in the mantle of science").

58. PCAST Report, *supra* note 3, at 46. For more on juror faith in forensic methods/experts, *see infra* Part III.C.

59. PCAST Report, *supra* note 3, at 59–63; *see also* Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 Sci. 892, 894 (2005) ("Unfortunately, forensic scientists often reject error rate estimates in favor of arguments that theirs is an error-free science."); NAS Report, *supra* note 4, at 128 (noting that forensics "developed heuristically . . . based on observation, experience, and reasoning without an underlying scientific theory, experiments designed to test the uncertainties and reliability of the method, or sufficient data that are collected and analyzed scientifically"); Jonathan J. Koehler et al., *The Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. Rev. 725, 730 (2011) ("Our central argument is that the pattern identification disciplines, and forensic science more generally, do not currently possess—and absolutely must develop—an adequate research culture."); Alex Biedermann & Christophe Champod, *Why the Post-Identification Era is Long Overdue: Commentary on the Current Controversy Over Forensic Feature Comparison as Applied to Forensic Firearms Examination*, 29 Int'l J. Evid. & Proof 30 (2024) ("Large parts of forensic science exhibit a longstanding tendency to focus on self-centered disciplinary discourses about the field's legitimacy and definition, with self-referential arguments about training and experience rather than classical research principles.").

consequences.[60] More often than not, however, sufficient data on accuracy simply does not exist.[61]

Filling such gaps around what we do not know about forensic accuracy is critical because "[t]he vast majority of jurors have no independent ability to interpret the probative value of results based on the detection, comparison, and frequency of scientific evidence."[62] And it is all the more so in the realm of firearms examination, which exceeds its forensic peers when it comes to unbridled subjectivity,[63] has failed to demonstrate that human judgment can accurately and consistently link fired ammunition back to a source gun,[64] and has yet to even prove that the "training and experience" of its participants

60. *See e.g., % Exonerations by Contributing Factor*, Nat'l Registry of Exonerations, *https://exonerationregistry.org/exonerations-contributing-factor* [https://perma.cc/XTW3-YEEZ] (last visited May 22, 2026) (Faulty or misleading forensic evidence is one of the leading causes of wrongful convictions, contributing to 29% of known exonerations overall and 50% of exonerations involving post-conviction DNA testing.); Simon A. Cole et al., *The Contribution of Forensic and Expert Evidence to DNA Exoneration Cases: An Interim Report*, Nat'l Registry Exonerations & Innocence Project (2022) (providing similar figures); Texas Forensic Science Commission, Forensic Bitemark Comparison Complaint Filed by National Innocence Project on Behalf of Steven Mark Chaney- Final Report 12, 15 (2016) (finding that "the overwhelming majority of existing research does not support the contention that bitemark comparison can be performed reliably and accurately" and therefore "recommending] that bitemark comparison not be admitted in criminal cases"); Kelly Sauerwein et al., *Bitemark Analysis*: A NIST Scientific Foundation Review, Nat'l Inst. Standards & Tech. Interagency 2 (2023) ("bitemark analysis lacks a sufficient scientific foundation . . . it has not been shown that defining characteristics of [bitemark] pattern[s] can be accurately analyzed to exclude or not exclude individuals as the source of a bitemark"); Jo-Anne Bright et al., *Internal Validation of STRmix™ – A Multi Laboratory Response to PCAST*, 34 Forensic Sci. Int'l: Genetics 11, (2018) (only 36% of study's participants correctly determined the number of contributors for five person mixtures, and none did so correctly for six person mixtures); John M. Butler et al., *NIST Interlaboratory Studies Involving DNA Mixtures (MIX05 and MIX13): Variation Observed and Lessons Learned*, 37 Forensic Sci. Int'l: Genetics 81, 87 (2018) (74 of 108 labs falsely included a contributor in the study's most complex mixture).

61. *See, e.g.*, NAS Report, *supra* note 4, at 8 ("A body of research is required to establish the limits and measures of performance and to address the impact of sources of variability and potential bias. Such research is sorely needed, but it seems to be lacking."); Saks & Koehler, *supra* note 59, at 895 ("[T]he time is ripe for the traditional forensic sciences to replace antiquated assumptions of uniqueness and perfection with a more defensible empirical and probabilistic foundation."); Koehler et al., *supra* note 59, at 736–37 ("[N]ot enough is yet known about a significant range of important questions. Consider, for example, the following: Precisely what are the capabilities and limitations of any particular pattern discipline? How often do pattern identification analysts make mistakes? When these errors occur, what causes them . . . the honest response to all of these questions is that we do not yet know.").

62. PCAST Report, *supra* note 3, at 45; *see* United States v. Yee, 134 F.R.D. 161, 181 (N.D. Oh. 1990) ("Without [empirical data] . . . the jury does not know whether the patterns are as common as pictures with two eyes, or as unique as the Mona Lisa.")

63. *See* Gutierrez, *supra* note 24, at 1546.

64. *See infra* Part I.A.

imparts expertise.[65] Proficiency tests[66] theoretically could establish reasonable estimates of discipline-wide error (while simultaneously giving courts a sense of the competence of specific examiners),[67] and they have been used by courts to do so.[68] Unfortunately, proficiency tests have historically failed to live up to this lofty purpose.

A. *The Uncharted Wilds of Firearms Examination*[69]

Firearms examination is a "feature comparison" methodology and a subspecies of toolmark examination.[70] It involves analysis of marks left behind on bullets or cartridge cases by tools in the form of various gun components (often lands and grooves of barrels for bullets and firing pin, breech face, and aperture shear marks for cartridge cases) when a firearm is discharged or cycled.[71] Specifically, examiners use a comparison microscope[72] to evaluate the similarities and differences found on two fired or cycled bullets or cartridge cases in terms of their "class characteristics" (design features chosen

65. *See* Gutierrez & Prokesch, *supra* note 13, at 1336, 1345 ("no validation effort that we are aware of has so much as established the existence of expertise in the comparison of firearms evidence"& "encourage[ing] the field of firearms examination to explore the role of expertise, as other fields such as latent print comparison, document examination, and facial recognition have attempted to do").
66. For a definition of proficiency testing *see infra* Part II.B.; Nat'l Comm'n on Forensic Sci., Views of the Commission, Proficiency Testing in Forensic Science 4–5 (Mar. 22, 2016); Jonathan J. Koehler, *Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter*, 59 Hastings L.J. 1077, 1091 (2008)
67. As discussed in far more detail later in this article, *see infra* Section III.C, this argument, at its core, is nothing new. *See e.g.*, Garrett & Mitchell, *supra* note 43, at 910 (arguing that "[a] proficiency-based approach to expert qualification would improve the quality of evidence at trial, and it would simplify judicial gatekeeping of expert evidence"); Jonothan J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences*, 12 L. Prob. & Risk 89, 89 (2013) ("But how accurate are forensic match reports? How can we provide legal decision makers with an empirically-based sense of the frequency with which various types of forensic testimony are wrong or misleading? The answer is to conduct methodologically rigorous, blind, external proficiency tests using realistic samples across the forensic sciences.")..
68. *See infra* Part II.B.
69. This part will provide only a brief overview of firearms examination methods and the criticisms directed at the field by outside scholars. For greater detail about firearms examination, *see* NRC Report, *supra* note 4, at 1–53; Robert M. Thompson, *Firearm Identification in the Forensic Laboratory*, Nat'l Dist. Atty. Ass'n (2010), and for a more in-depth history of criticisms leveled at the discipline, *see* Gutierrez, *supra* note 24, at 1541–53; Defendant's Post-Frye/Rule 403 Hearing Brief Against the Admissibility of Firearms Examination Evidence, *Illinois v. Winfield*, No. 15CR14066–01 (Cir. Ct. Cook Cnty. Sept. 14, 2022).
70. PCAST Report, *supra* note 3, at 23, 104 ([S]uch methods "aim to determine whether an evidentiary sample . . . is or is not associated with a potential source sample . . . based on the presence of similar patterns, impressions, features, or characteristics in the sample and the source.").
71. *Id.* at 104; NAS Report, *supra* note 4, at 150–51.
72. *See generally* Thompson, *supra* note 69, at 8 ("The principles and the primary tools used in the science have changed very little during this time. The comparison microscope, the primary tool used by the profession, has not changed in its basic design for almost 80 years.").

by the manufacturer like caliber, number and direction of twist for lands and grooves, and overall shape of firing pin marks)[73] and "individual characteristics" (microscopic scratches and impressions resulting from purportedly random irregularities of gun component working surfaces).[74] Because they are common to large numbers of guns, class characteristics only allow examiners to "group guns into those that could, versus could not, have fired a particular bullet or cartridge case," but do not allow them to single out any one particular gun as the weapon responsible for discharging or cycling a particular munition.[75] In contrast, when examiners encounter "sufficient agreement" of individual characteristics on two bullets or cartridge cases, they provide far stronger "opinions of common origin."[76] Technically, those opinions break down into three categories: (1) identification opinions which state that the bullets or cartridge cases were fired by the same gun; (2) elimination conclusions which state that the bullets and cartridge cases were fired by different guns; and (3) inconclusive opinions which state that an examiner did not observe enough similarity or dissimilarity to justify either of the two previous opinions.[77] But in practice, examiners are loath (and sometimes outright prohibited by laboratory policy) from reaching elimination conclusions based on individual characteristics.[78]

73. Ass'n of Firearms & Toolmark Exam'r's, Glossary 39 (6th ed. 2013).
74. *Id.* at 66.
75. Gutierrez, *supra* note 24, at 1541–42.
76. "Sufficient agreement" is a term of art set by a trade organization called the Association of Firearms and Toolmark Examiners ("AFTE") and meaning "when the agreement in individual characteristics 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"—they will provide far stronger "opinions of common origin." *Theory of Identification: Revisited*, *supra* note 7, at 287. It has been heavily criticized as tautological. PCAST Report, *supra* note 3, at 60, 104 (saying of the AFTE Theory "the stated method is circular. It declares that an examiner may state that two toolmarks have a 'common origin' when their features are in 'sufficient agreement.' It then defines 'sufficient agreement' as occurring when the examiner considers it a 'practical impossibility' that the toolmarks have different origins."). Multiple courts have accepted this criticism as sound. *See, e.g.*, United States v. Shipp, 422 F. Supp. 3d 762, 779 (E.D.N.Y. 2019); United States v. Taylor, 663 F. Supp. 2d 1170, 1177 (D.N.M. 2009).
77. *See Range of Conclusions*, Ass'n of Firearm Toolmark Exam'rs, https://afte.org/about-afte/afte-range-of-conclusions [https://perma.cc/2K9C-LF3U] (The inconclusive bin is sometimes split, creating a five-category system, to account for examiners leaning in the direction of an identification or elimination.).
78. *See, e.g.*, Maneka Sinha & Richard E. Gutierrez, *Signal Detection Theory Fails to Account for Real-World Consequences of Inconclusive Decisions*, 21 L., Prob., & Risk 131, 132 (2022) ("[W]hether due to adversarial allegiance bias, training and standards that focus more on identifications than exclusions, lab protocols that disallow exclusion decisions unless class characteristics disagree or outside of 'exceptional' circumstances, or for other reasons, examiners far more readily default to inconclusive decisions where the data supports exclusion than where the data supports identification . . . what we call the *Biased Examiner*."); Laura Knowles et al., *The Validation of 3D Virtual Comparison Microscopy (VCM) in the Comparison of Expended Cartridge Cases*, 67 J. Forensic Sci. 516, 518 (2021) ("[E]liminations based on an apparent difference in individual characteristics are rarely, if ever, made when comparing toolmarks between only two items."); David P. Baldwin et al., *A*

Although the Association of Firearms and Tool Mark Examiners (AFTE)[79] supports these conclusions by claiming that manufacturing irregularities and use of a firearm result in "unique surface contours," it also admits that "[c]urrently the interpretation of individualization/identification is subjective in nature . . . and based on the examiner's training and experience."[80] In other words, however rare[81] the features observed by practitioners, how well examiners can actually perceive and interpret those features to arrive at source conclusions remains key to vetting the field's legitimacy.[82] As PCAST eloquently puts it:

> 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

---

Study of Examiner Accuracy in Cartridge Case Comparisons. Part 1: Examiner Error Rates, 349 Forensic Sci. Int'l, August 2023, at 7 (noting a "large number of examiners who through training or agency policy choose only to report Inconclusive conclusions in the absence of an Identification or class characteristics differences that would lead to an Elimination."); 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. Appl. Res. Memory & Cognition 147, 148 (2023) ("[A] forensic laboratory might adopt a policy that prohibits its examiners from rendering elimination reports.").

79. AFTE is the main trade association for firearms examiners and promulgates multiple widely-relied-upon resources for practitioners. *See* Ass'n of Firearm & Toolmark Exam'rs, *https://afte.org/about-afte/what-is-afte/* [https://perma.cc/T3JN-BBH4]; *see also* Hinton v. Alabama, 571 U.S. 263, 270 (2014).

80. Comm. for the Advancement of the Sci. of Firearm & Toolmark Identification, *supra* note 7, at 287.

81. Firearms examination currently lacks robust empirical data on the rarity of features, with algorithms that could produce objective metrics of similarity still under development. *See* Heike Hofmann et al., *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 L. Probab. & Risk 317, 344 (2020); *see also* PCAST Report, *supra* note 3, at 114; Eric Hare, Heike Hofmann & Alicia Carriquiry, *Automatic Matching of Bullet Land Impressions*, 11 Ann. Appl. Stat. 2332, 2354 (2017). But multiple signs point to AFTE's claims being overblown. Studies show that munitions fired by different weapons can randomly display levels of similarity exceeding that created by multiple firings of the same gun. *See* Alfred A. Biasotti, *A Statistical Study of the Individual Characteristics of Fired Bullets*, 4 J. Forensic Sci. 34, 38–40 (1959); 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); William Tobin & Peter Blau, *Hypothesis Testing of the Critical Underlying Premise of Discernible Uniqueness in Firearms-Toolmark Forensic Practice*, 53 Jurimetrics 121, 136 (2013). And examiners have long been aware of a third type of marking, "subclass characteristics," that mimic individual characteristics but actually appear across gun components made back-to-back in the same batch. *See* Glossary, *supra* note 73, at 122; Gene C. Rivera, S*ubclass Characteristics in Smith & Wesson SW40VE Sigma Pistols*, 39 AFTE J. 247 (2007); 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).

82. Biedermann & Champod, *supra* note 59, at 34–35.

> *detect* relevant differences in features to reliably identify whether they share a common source.[83]

Because "[e]xpertise does not guarantee reliability or translate into a low risk of error,"[84] firearms examination's dependence on the subjective human judgment of purported experts therefore "require[s] particularly careful scrutiny."[85] In other words, we need rigorous validation[86] testing aimed at establishing the method's "capabilities and limitations," its ability to "accurately and reliably answer" whether a bullet or cartridge case was fired and/or cycled through a particular gun.[87] Reliability[88] breaks down into repeatability (the probability that "an examiner obtains the same result, when analyzing samples from the same sources") and reproducibility (the probability that "different examiners obtain the same result, when analyzing the same samples").[89] But because examiners might be consistently wrong, these metrics alone cannot paint the full picture of a method's validity.[90] We must also know how often examiners get the correct answer, based on their sensitivity, specificity, false positive rate, and false negative rate.[91] Beyond that, the empirical tests conducted to establish

83. PCAST Report, *supra* note 3, at 62.
84. Jonothan J. Koehler, *Fingerprint Error Rates & Proficiency Tests: What They are & Why They Matter*, 59 Hastings L.J. 1077, 1087 (2008) (along similar lines, success in casework)
85. PCAST Report, *supra* note 3, at 5.
86. *See* Nat'l Comm'n on Forensic Sci., Views of the Commission: Validation of Forensic Science Methodology 2 (2016) (defining validation as "a process by which a procedure is evaluated to determine its efficacy and reliability for forensic analysis").
87. Nat'l Comm'n on Forensic Sci., Views of the Commission: Technical Merit Evaluation of Forensic Science Methods and Practices 2 (2016); *see also* Thomas D. Albright, *How to Make Better Forensic Decisions*, 119 Proc. Nat'l Acad. Sci. U.S.A., no. 38, 2022, at 9; Hofmann et al., *supra* note 81, at 318–19; Itiel E. Dror & Nicholas Scurich, *(Mis)use of Scientific Measurements in Forensic Science*, 2 Forensic Sci. Int'l Synergy 333 (2020).
88. Scientists use this term interchangeably with consistency. *See* Hum. Factors Task Grp., Org. of Sci. Area Comms. for Forensic Sci., Human Factors in Validation & Performance Testing of Forensic Science, OSAC Technical Series 0004 3 (2020) [hereinafter OSAC Human Factors Report].
89. PCAST Report, *supra* note 3, at 47.
90. OSAC Human Factors Report, *supra* note 88, at 6–7.
91. *Id.*; PCAST Report, *supra* note 3, at 5. Specificity and sensitivity are rates of correct conclusions: specificity is "the probability that examiners will deem two items to be from a different source when they are actually from different sources," and sensitivity is "the probability that examiners will deem two items to be from the same source when they are from the same source." OSAC Human Factors Report, *supra* note 88, at 4–5. Errors are most commonly measured by the false positive rate ("the probability that the method declares a proposed identification between samples that actually come from *different* sources") and the false negative rate (the probability that the method excludes or eliminates even though samples come from the *same* source). PCAST Report, *supra* note 3, at 50. Though all these rates are critical poor performance on different source samples (weak specificity / a high false positive rate) most directly imperils innocent defendants and so will be the focus on this paper. *See id.*; Simon A. Cole & Barry Sheck, *Fingerprints and Miscarriages of Justice: 'Other' Types of Error and A Post- Conviction Right to Database Searching*, 81 Alb. L. Rev. 807, 810, 819 (2018); *Abruquah*, 296 A.3d at 989 n.21; Sinha & Gutierrez, *supra* note 78, at 133. For

these rates of accuracy and consistency must satisfy widely accepted scientific standards: They should involve sufficiently large samples of examiners and comparison items; reflect the full range of casework challenges; transparently subject their data to independent scrutiny; incorporate individuals without a stake in the success of the method as researchers; and (preferably) be blind, i.e., occur without the participant knowing they are being tested.[92]

While firearms examination has faced a host of other criticisms—directed at the indefensibility of uniqueness claims,[93] the absurdity of asserting certainty in conclusions,[94] the lack of well-defined processes or thresholds,[95] and even its inconsistency with human cognitive abilities[96]—the field falls shortest when it comes to establishing these useful estimates of accuracy and consistency with sound scientific studies. Since the late 1990s, practitioners, members of law enforcement,[97] and academic researchers[98] have conducted close to thirty

more on measuring accuracy, *see generally* Koehler, *supra* note 84, at 1081–86.

92. *See, e.g.*, PCAST Report, *supra* note 3, at 52; Cueller et al., *supra* note 30, at 143–45; OSAC Human Factors Report, *supra* note 88, at 8–24; Jonothan J. Koehler, *Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences*, 49 Ariz. St. L.J. 1369, 1398–1406 (2017); Itiel E. Dror, *The Error in "Error Rate": Why Error Rates Are So Needed, Yet So Elusive*, 65 J. Forensic Sci. 1034 (2020); Nicholas Scurich et al., *Scientific Guidelines for Evaluating the Validity of Forensic Feature-Comparison Methods*, 120 Proc. Nat'l Acad. Sci. U.S.A. e2301843120, 4–7 (2023); Hal R. Arkes & Jonathan J. Koehler, *Inconclusives and Error Rates in Forensic Science: a Signal Detection Theory Approach*, 20 L. Probab. & Risk 153, 160 (2021); Nicholas Scurich et al., *The Hawthorne Effect in Studies of Firearm and Toolmark Examiners*, 70 J. Forensic Sci. 1329, 1336 (2025); John M. Butler et al., DNA Mixture Interpretation: A NIST Scientific Foundation Review, NISTIR 8351-draft, at 60–61 (Nat'l Inst. Standards & Tech. 2021). Success is casework is no substitute for such rigorous testing. *See* Koehler, *supra* note 84, at 1086–87; Simon A. Cole, *Can Casework Validate Forensic Techniques?*, 46 Jurimetrics 117, 126–27 (2006).

93. *See, e.g.*, Mark Page et al., *Uniqueness in the Forensic Identification Sciences—Fact or Fiction?*, 206 Forensic Sci. Int'l 12, 14–15 (2011) ("The concept of 'uniqueness' has more the qualities of a cultural meme than a scientific fact"); PCAST Report, *supra* note 3, at 61–62; NRC Report, *supra* note 4, at 82.

94. *See, e.g.*, PCAST Report, *supra* note 3, at 19; Simon A. Cole, *Individualization is Dead, Long Live Individualization! Reforms of Reporting Practices for Fingerprint Analysis in the United States*, 13 L. Probab. & Risk 117 (2014); William Tobin & Peter Blau, *supra* note 81, at 131. Indeed, even writers otherwise supportive of firearms examination have decried its source attribution and certainty statements as overblown and scientifically indefensible. *See, e.g.*, Biedermann & Champod, *supra* note 59, at 13 ("[E]xperts should abandon the practice of providing source attribution conclusions . . . [i]f professional forensic organizations were genuinely committed to science, they would adopt this proposed change in reporting practice *today.*"); Nada Aggadi et al., *Quantifying the Strength of Firearms Comparisons Based on Error Rate Studies*, 70 J. Forensic Sci. 84 (2025) (describing firearms examiner claims as far stronger than existing studies support).

95. *See, e.g.*, NAS Report, *supra* note 4, at 153–55.

96. *See* Scurich et al., *supra* note 92, at 4.

97. *See, e.g.*, M. A. Keisler et al., *Isolated Pairs Research Study*, 50 AFTE J. 56 (2018); David J. Brundage, *The Identification of Consecutively Rifled Gun Barrels*, 30 AFTE J. 438 (1998).

98. *See, e.g.*, Max Guyll et al., *Validity of Forensic Cartridge-Case Comparisons*, 120 Proc. Nat'l Acad. Sci. U.S.A., no. 20, 2023.

studies aimed at estimating these critical performance metrics.[99] But as one group of researchers bluntly concludes, "[s]ince . . . [these] validation studies . . . have . . . fundamental flaws . . . the testing they provide is inadequate. This means that the reliability of firearms examination is unknown, posing a significant risk of contributing to incorrect determinations of guilt or innocence."[100] Early studies in the field[101] used "set-based" designs that actually preclude any false positive rate calculations and (like a Sudoku puzzle) would "inflate examiners' performance by allowing them to take advantage of internal dependencies."[102] With one exception,[103] none have been blind, even though examiners do appear to behave differently (and in ways that artificially avoid errors) when they know they are being tested.[104] Across the board, these studies suffer from basic data collection issues, like inadequate and unrepresentative sampling, missingness, and poor accounting for uncertainty.[105] And few have made any legitimate efforts to cover the full range of difficulty anticipated in casework by seeking out and including challenging comparisons.[106] Indeed, regarding the last of those, one group of authors outright admitted to avoiding difficult elimination comparisons,[107] another[108] used samples that posed so little difficulty that lay people could and did achieve false positive rates indistinguishable from professional examiners,[109] and a third[110] failed to include elimination samples with "similarities, features that might kind of dupe or trick an examiner into making a false positive."[111]

99. *See* Cuellar et al., *supra* note 30, at 14–17; U.S. Dep't of Justice, Statement on the PCAST Report: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, at 23–26 (2021), https://www.justice.gov/opa/pr/justice-department-publishes-statement-2016-presidents-council-advisors-science-and [https://perma.cc/YH2N-LBK9] [hereinafter DOJ Statement].
100. Cuellar et al., *supra* note 30, at 12.
101. *See, e.g.*, Brundage, *supra* note 97; B. Mayland, B. & C. Tucker, *Validation of Obturation Marks in Consecutively Reamed Chambers*, 44 AFTE J. 167 (2012).
102. *See e.g.*, PCAST Report, *supra* note 3, at 106, 109; Cuellar et al., *supra* note 30, at 15; Hofmann et al., *supra* note 81, at 332–33.
103. Maddison Neuman et al., *Blind Testing in Firearms: Preliminary Results from a Blind Quality Control Program*, 67 J. Forensic Sci. 964 (2022).
104. *See* Scurich et al., *supra* note 92, at 1336–37 (noting that their findings that examiners more often respond inconclusive when they are aware of a test "rebut claims that there is no evidence that firearms examiners behave differently while being tested").
105. *See, e.g.*, Cueller et al., *supra* note 30, at 3–12; Khori Khan & Alicia Carriquiry, *Shining a Light on Forensic Black-Box Studies*, 10 Stat. & Pub. Pol'y 2216748 (2023).
106. Gutierrez, *supra* note 24, at 1561–73.
107. Knowles et al., *supra* note 78, at 517.
108. Pierre Duez et al., *Development and Validation of a Virtual Examination Tool for Firearm Forensics*, 63 J. Forensic Sci. 1069 (2018).
109. Gutierrez & Prokesch, *supra* note 13, at 1341–42.
110. Chapnick *et al., Results of the 3d Virtual Comparison Microscopy Error Rate (VCMER) Study for firearm forensics*, 66 J. Forensic Sci. 557 (2021).
111. Richard E. Gutierrez & Emily J. Prokesch, *Authors' Response to Weller et al. Commentary on: Gutierrez RE, Prokesch EJ. The False Promise of Firearms Examination Validation Studies: Lay Controls, Simplistic Comparisons, and the Failure to Soundly Measure Misidentification Rates.*, 70 J. Forensic Sci. 405, 407 (2024) (citation omitted).

Critically, "since the studies all share the[se] fundamental flaws . . . they cannot produce reliable conclusions . . . like a set of bagels that align such that there is a wide hole . . . that pierces through all the bagels at once, no matter how many you stack together," the sheer number of validation efforts cannot save the field.[112] But perhaps worse, even putting aside their myriad deficiencies, existing efforts to measure the accuracy and consistency of firearms examination paint a less-than-rosy picture of the discipline's trustworthiness. Examiners disagree with one another (even when comparing identical evidence) almost as often as (sometimes more than) they manage to reach the same conclusion.[113] Indeed, scholars have called such performance (with almost comical scientific civility) "rather weak."[114] False positive rates have repeatedly climbed into the double digits.[115] And, across study after study, examiners display a systematic bias against elimination conclusions, and thus against criminal defendants.[116]

But the truly damning reality is the breadth of what we do not know. Accuracy rates clearly appear to be "associated with a variety of factors," such that one cannot port performance from one set of guns or one set of examiners over to any other.[117] Test examiners on Glock firearms (long considered among the most simplistic comparisons)[118] and you may get a 0% false positive rate.[119] But present examiners with subclass characteristics (or even just Glock comparisons in which researchers sought out challenging sets of fired cartridge cases) and false positive rates climb to 38.9% or 10.8% respectively.[120] Check

112. Cueller *et al.*, *supra* note 30, at 13.
113. *See* R. Austin Hicklin et al., *Accuracy and Reproducibility of Bullet Comparison Decisions by Forensic Examiners*, 365 Forensic Sci. Int'l 112287, 11 (2024) (reporting 56.9% reproducibility on bullet comparisons); Keith L. Monson, Eric D. Smith & Eugene M. Peters, *Repeatability and Reproducibility of Comparison Decisions by Firearms Examiners*, 68 J. Forensic Sci. 1721, 1729 (2023) (reproducibility for bullet comparison eliminations of 31.62% and for cartridge case comparison eliminations of 41.29%).
114. Allan 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).
115. *Supra* note 32.
116. *See* Smith & Wells, *supra* note 78, at 152 ("[A] huge bias against reporting exculpatory evidence by hiding the exculpatory evidence in an inconclusive category."); Sinha & Gutierrez, *supra* note 78, at 132 (documenting the bias against exculpatory evidence across multiple error rate studies on firearms examination).
117. Hicklin et al., *supra* note 113, at 13; *see also* Bullet Black Box Working Grp. , NIST GCR 25–068: Implications of the Bullet Black Box Study 3 (2025) (Natl Inst. Standards & Tech. 2025) ("[R]ates and distributions of decisions [in firearms examination] are associated with a variety of factors . . . the wide variation among all these factors leads us to caution that citing overall rates without differentiating among these factors is likely to be misleading.").
118. *See* 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 Sci. 170, 172 (2016); Stephen G. Bunch & Douglas P. Murphy, *A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases*, 35 AFTE J. 201, 202 (2003); Angela Stroman, *Empirically Determined Frequency of Error in Cartridge Case Examinations Using a Declared Double-Blind Format*, 46 AFTE J. 157, 158 (2014).
119. *See* Keisler et al., *supra* note 97, at 57.
120. Pauw-Vugts et al., *supra* note 32, at 125 (subclass comparisons); Mattijssen et al.,

the individualized performance of examiners in a study and you will find loads who commit no errors, but so too will you find examiners who essentially only err.[121] Tally up the total variables explored across studies capable of measuring a false positive rate and you will see glaringly few firearm makes and models represented,[122] and no answer as to why some examiners are better than others, no way to predict whether you are staring down the barrel at a skilled expert or a snake oil salesman.[123] Although firearms examiners have, almost paradoxically,

supra note 32, at 7 (challenging Glock comparisons); Gutierrez, *supra* note 24, at 1569 & 1571 (further explaining studies and comparison samples); *see also* Guyll et al., *supra* note 98, at 8 (noting performance differences on Berretta versus HiPoint comparisons and concluding that "highlight the importance of considering whether a technique can be accurately characterized under all conditions present in the field by a single set of performance values."); Bullet Black Box Working Grp., *supra* note 117, at 4–5 ("decision rates varied significantly by firearm make/model and by ammunition type" & "lower quality bullets (i.e., bullets with more damage) tend to have higher rates of inconclusive responses").

121. Bullet Black box Working Grp., *supra* note 117, at 9–10 ("Decision rates varied notably among the participants in the study . . . [a] few participants made a disproportionate number of errors . . . [t]he fact that any practicing [examiners] had such high error rates is an obvious concern."); Gutierrez, *supra* note 31, at 4 (noting that in one study approximately 2% of examiners committed false positives, around 18% of examiners never reached an elimination, and those groups combined committed false positives 28% of the time they reached conclusive source determinations); Gutierrez & Addyman, *supra* note 31, at 1097 (in one study approximately 6% of examiners committed false positives on bullet comparisons with a false positive rate of 21% and 10% of examiners committed false positives on cartridge case comparisons for a false positive rate of 17%).

122. Gutierrez, *supra* note 24, at 1561 (cataloging that the studies that have produced false positive rates favored by firearms examiners tested just one caliber, two manufacturers, and 38 guns for bullets, and two calibers, six manufacturers, and 98 guns for cartridge cases). One additional study has emerged since, though it limits itself as an "exploratory, hypothesis generating study, and thus should likely not be added to these values. Hicklin et al., *supra* note 113, at 12. If considered it would bring the bullet totals up to four calibers, six manufacturers, and 92 guns. *See id.* at 9 (2025). This coverage compares unfavorably, to say the least, with the expectations of experts. *See Kimberley*, No. 22-cr-02932–01, at 6 (noting testimony by Dr. Maria Cuellar estimating that a robust sample of participants and comparison items would require "at least 'hundreds' of examiners examining projectiles from 800 guns.").

123. *See, e.g.*, *Kimberley*, No. 22-cr-02932–01,.at 20 (noting testimony of Dr. William Thompson that "some examiners are better than others for reasons that are not currently understood. So, inexplicably, some examiners are getting a lot more correct than others."); Gutierrez, *supra* note 31, at 3 (findings regarding poor performing examiners "leave triers of fact in the criminal legal system in the unenviable position of being required to vet the qualifications of firearms examination witnesses, knowing performance varies widely and into double- digit rates of misidentification among practitioners, and yet bereft of tools to sort experts from charlatans."); Gutierrez & Addyman, *supra* note 31, at 1100 ("realities [regarding poor performing examiners] leave the court system in a difficult bind . . . they have precious little (if any) means at their disposal to assess the . . . important question of whether a given examiner is qualified to offer expert testimony: none of training, experience, or accreditation appear to predict accuracy and empirical measures are lacking at the level of individual examiners.").

treated this resounding silence as a defense,[124] what it truly means is that the "trier of fact . . . lacks information that would allow them to determine where on the error spectrum the testifying examiner falls."[125] The present record underlying firearms examination, then, demands speculation and the dreaded *ipse dixit*[126] far more than it could hope to instill anything resembling confidence. Powerful before juries and popular as it may remain with prosecutors,[127] it should therefore come as little surprise that courts have begun to question their longstanding guardianship of the field[128] and that it has found no real allies amidst the credentialed and the disinterested.[129]

124. Essentially, examiners have used the field's variability to dismiss damaging false positive rates as insufficiently representative when they arise. *See, e.g.*, Carroll et al., *supra* note 28, at 142 ("As demonstrated by the differences in error rates between different demographic groups . . .the PTRC strongly recommends against calculating an overall false positive error rate from this test as being reflective of the discipline as a whole or reflective of any individual examiner.").
125. *State v. Adams*, 572 P.3d 291, 312–13; *supra* note 22.
126. *See, e.g.*, Gen. Elec. Co.v. Joiner, 522 U.S. 136, 146 (1997) ("[N]othing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert. A court may conclude that there is simply too great an analytical gap between the data and the opinion proffered.").
127. *See* Garrett et al., *supra* note 36, at 224, 241–42 (noting how persuasive jurors find firearms examination evidence and also commenting on its popularity: "Such testimony continues to be in great demand, given the degree of firearm-related violence in the United States, with more than 100,000 requests for bullet or cartridge case comparisons each year."); Raymond Valerio & Nelson Bunn, *Firearm Forensics Has Proven Reliable in the Courtroom. And in the Lab*, Sci. Am. (Nov. 27, 2023) ("As members of the National District Attorneys Association, we advocate for the use of reliable forensics to exonerate the innocent and inculpate the guilty. NDAA prosecutors, who are the "boots on the ground" in courtrooms throughout this country, know from experience that firearms identification evidence is scientifically sound and withstands rigorous testing in the crucible of the courtroom."), *https://www.scientificamerican.com/article/firearm-forensics-has-proven-reliable-in-the-courtroom-and-in-the-lab* [https://perma.cc/ZF9L-TKQZ]; Colonel (Ret.) Jim Agar, *The Admissibility of Firearms & Toolmarks Expert Testimony in the Shadow of PCAST*, 74 Baylor L.R. 93, 195 (2022) (discussing homicides in the United States and the common use of guns to commit those crimes before arguing: "Given this fundamental threat to public health and safety, it seems absurd that some courts would attempt to diminish or exclude firearms identification expert testimony, which is often central to the prosecution of such cases").
128. *See supra* notes 38–39; *Kimberley*, No. 22-CR-02932–01, at 43–44, 46,50–51 ("it is abundantly clear that errors . . . are wildly varying and extremely dependent on testing conditions;" expressing "concern about whether accuracy in firearms identification is truly quantified;" "there is a large disparity in error rates . . . the significance of stated error rates might not reflect reality").
129. *Supra* note 35. This author is aware of only one group of purportedly disinterested academics who have described firearms examination as valid. *See* Guyll *et al.*, *supra* note 98, at 7 ("[C]artridge-case comparison is a highly valid forensic technique."). Two of those authors, however, have also characterized conclusions in the field as systemically and consciously biased against defendants. *See* Smith & Wells, *supra* note 78, at 152. And their original study and characterization of firearms examination as valid has met with substantial scholarly opprobrium. *See* Gutierrez, *supra* note 24, at 6 (calling comments on validity nothing other than "hyperbole and incaution");

B. *Proficiency Tests & Consumer Satisfaction, A Recipe for Meaningless Results*

Given the significant empirical gaps currently endemic to firearms examination, well-designed proficiency tests[130] could play a prominent role in judicial and juror evaluations of the discipline (as well as of the qualifications of its individual practitioners). These tests involve assessing examiner "performance against preestablished criteria."[131] Put another way, they help ensure that examiners can and will reach appropriate conclusions about the source of items of interest to an investigation (in a casework scenario) by providing "samples whose sources are known to the proficiency test administrator but unknown to the examinee" (in a controlled testing scenario) and seeing if laboratory personnel can reach the correct conclusion.[132] In the case of external proficiency tests like those offered by CTS,[133] a company first develops comparison sets of known samples (fingerprints, bullets, or anything else forensic scientists might analyze) that examiners must compare against questioned samples to identify or exclude a common origin.[134] Before releasing a test, the company then "helps ensure test validity" by distributing the samples for review by discipline experts.[135] Laboratories (or individuals) order and receive the test, complete it according to their own varying protocols, and return their answers to the

Rosenblum *et al.*, *Misuse of Statistical Method Results in Highly Biased Interpretation of Forensic Evidence in Guyll et al. (2023)*, 23 L. Prob. & Risk. 1, 5 (2024) (explaining that Guyll and his coauthors produced "a highly misleading posterior probability that the reference gun fired the crime scene cartridge case, given a firearms examiner's 'identification' decision"); Rosenblum *et al.*, *Incorrect Statistical Reasoning in Guyll et al. Leads to Biased Claims About Strength of Forensic Evidence*, 121 Proc. Nat'l Academy Scis. e2315431121, at 1 (2024) (describing Guyll and coauthors as making "an error in statistical reasoning that could lead judges and jurors in criminal trials to grossly misinterpret forensic evidence"); Alicia Carriquiry & Danica M. Ommen, *What Are "Fair" Prior Odds of Guilt in Criminal Cases?*, 121 Proc. Nat'l Academy Scis. no. 45, 2024), at 1 (critiquing the statistical reasoning of Guyll and his coauthors and concluding that "[t]hey have not shown that firearms examination is 'valid'"). Indeed, even other academics hired by the prosecution to defend the field's admissibility have refused to fully endorse the field's subjective conclusions as valid. *See*, *e.g.*, Alex Biedermann et al., *Forensic Feature-Comparison as Applied to Firearms Examination: Evidential Value of Findings and Expert Performance Characteristics: Reply to the Affidavit by Vanderplas et al. (2022)*, United States v. Sutton, No. 2018 CF1 009709, at 6, 6–18 (D.C. Super. Ct. Apr. 28, 2022) (calling the question of validity "elusive" and as having "no easy answer" and saying "the performance of human examiners in conducting feature-comparisons, is more nuanced to tackle because of the debates over what, if anything, one can legitimately conclude from existing error rate studies").

130. For an in-depth treatment of appropriate design for proficiency tests, *see generally* Koehler, *supra* note 84.
131. Nat'l Comm'n on Forensic Sci., *supra* note 66, at 4–5.
132. Koehler, *supra* note 66, at 1091.
133. Laboratories may also create and distribute tests internally. *See, e.g.*, Nat'l Comm'n on Forensic Sci., *supra* note 66, at 5.
134. *See* Collaborative Testing Servs., Inc., *About CTS: How it Works*, *https://cts-forensics.com/index-forensics-testing.php#row-2* [https://perma.cc/6EQC-4URM] (last visited May 22, 2026).
135. *See id.*

company.[136] CTS specifically then releases results in two formats: "Summary Reports" and "Individual Reports."[137] Summary Reports from the last five years are available to the public on CTS's website, but they anonymize participants, identifying them only by a "Webcode" which "is randomly assigned and changes with every report."[138] In contrast, Individual Reports are not anonymous but are sent only to participating laboratories.[139]

These proficiency tests fill critical needs internal to laboratories—"allowing for training and testing personnel, improving laboratory practices and procedures, and identifying future needs"[140]—and fulfill accreditation requirements.[141]

136. CTS does not monitor test takers or mandate any specific procedure for completing its tests beyond imposing a due date. *Id.*; *see* Koehler, *supra* note 66, at 1091 ("[I]t is not clear whether examiners work by themselves, in groups, of with the assistance of supervisors."); Brett O. Gardner, *et al.*,, *Latent Print Proficiency Testing: An Examination of Test Respondents, Test-Taking Procedures, and Test Characteristics*, 65 J. FORENSIC SCI. 450, 450,453 (2020) (noting wide variation in the procedural approach taken by labs to CTS proficiency tests); Simon A. Cole, *More Than Zero: Accounting for Error in Latent Fingerprint Identifications*, 95 J. CRIM. L. & CRIMINOLOGY 985, 1029 (2005) (describing tests as "conducted by mail under unproctored, untimed conditions. It is not known whether the tests were completed by individual examiners or by committee."); Amanda S. Luby & Joseph B. Kadane, *Proficiency Testing of Fingerprint Examiners with Bayesian Item Response Theory*, 17 L., PROB. & RISK 111 (2018) ("[S]ince the testing environment is not controlled, it is unclear whether each response corresponds to an individual examiner or represents the consensus answer of a group of examiners working together on the exam."); *see also* Brendan Max et al. , *Assessing Latent Print Proficiency Tests: Lofty Aims, Straightforward Samples, & the Implications of Nonexpert Performance*, 69 J. FORENSIC ID. 281, 284–86 (2019) (discussing participation in a CTS proficiency test).

137. COLLABORATIVE TESTING SERVS., INC., *supra* note 134.

138. *Id.*

139. *Id.*

140. Koehler, *supra* note 66, at 1091; *see* NAT'L COMM'N ON FORENSIC SCI., *supra* note 66, at 3 (listing multiple purposes for labs including "evaluation of the performance of laboratories for specific tests or measurements and monitoring laboratories' continuing performance" & "identification of problems in laboratories and initiation of actions for improvement"); Koehler, *supra* note 84, at 1400 (explaining that many "proficiency tests serve purposes that are primarily internal . . . are designed to identify the strengths and weaknesses in procedures and personnel for the internal purpose of improving forensic science work . . . [and] addressing such questions as, 'Are the training programs sufficient?' 'Are the laboratory protocols clear?' and 'Are laboratory personnel able to follow the protocols to generate a result that competent others obtain?'" (emphasis omitted)); Cole, *supra* note 136, at 1028 ("The purpose of proficiency tests is to measure the competence of individual laboratories or techniques"); Sharon Kelley et al., *How Do Latent Print Examiners Perceive Proficiency Testing? An Analysis of Examiner Perceptions, Performance, and Print Quality*, 60 SCI. & JUST. 120, 120 (2020) ("Proficiency testing, often part of laboratories' quality assurance processes, can help train incoming forensic analysts and establish base levels of competency within and across laboratories. Routine proficiency testing can also identify problematic laboratory procedures and ultimately improve analyst practice.").

141. *See, e.g.*, Garrett & Mitchell, *supra* note 43, at 916, 922–23 ("In the area of forensic science, all accredited crime laboratories must conduct proficiency testing annually across the different disciplines that they employ, whether it is DNA testing, fingerprint testing, ballistics, toolmark identification, or some other discipline."); NAS REPORT, *supra* note 4, at 169–171 ("Proficiency testing requirements are agreed upon by the

One group of authors has even written at length about how errors on such tests "during training may also have positive consequences in that they may benefit the learning process and result in fewer errors during what is termed 'the test that counts'—which, in the case of forensic examiners, includes conclusions, reporting, and testimony in casework."[142] But they could also serve purposes external to laboratories and essential to the criminal legal system: "[p]roperly designed, proficiency tests may also provide a reasonable estimate of the rate at which . . . " different types of error may occur in casework,[143] as well as allow courts to engage in empirical and defensible evaluations of the competence of individual examiners who appear as witnesses (to see, in other words, where along the error spectrum a given examiner falls).[144] Indeed, multiple blue-ribbon panels and standard-setting organizations have called for mandatory and rigorous proficiency testing.[145] And forensic practitioners have not shied away from deploying proficiency test results to these ends,[146] famously (as just one example) including them as part of a successful effort to seek

predominant accrediting organization, the American Society of Crime Laboratory Directors-Laboratory Accreditation Board (ASCLD/LAB), which requires testing (internal or external) once per calendar year".). Proficiency tests are required by statute for use of DNA databases and the FBI. *See* 42 U.S.C. §§ 14132(b)(2), 14133(a)(1)(A).

142. Heidi Eldridge et al., *The Benefits of Errors During Training*, 4 Forensic Sci. Int'l Synergy, 2022, at 1.

143. Koehler, *supra* note 66, at 1091; *see also* Cole, *supra* note 136, at 1028–29 ("[T]heir intent is not to generate an estimate of the accuracy of latent print identification. Nonetheless, proficiency tests are simulations; the correct answer is known to the testmaker, and it is possible to measure the number of correct and incorrect responses. It, therefore, may be possible to infer something about the accuracy of forensic fingerprint identification from existing proficiency tests data."); Nat'l Rsch. Council, Fed. Jud. Ctr., Reference Manual on Scientific Evidence 28 (3d ed. 2011) ("In theory, proficiency tests could determine whether well-trained experts in those fields can reach results with low error rates. In practice, however, there are numerous obstacles to such tests."); William Thompson et al., *Forensic Science Assessments, A Quality & Gap Analysis: Latent Print Examination*, Am. Assoc. for the Advancement of Science 53 (2017) ("While we agree that these proficiency tests are not well designed for error-rate estimation, we believe the results of these tests nevertheless provide useful information about the potential for error. If nothing else, they refute claims that errors are extraordinarily unlikely events.").

144. *See, e.g.,* Thompson et al., *supra* note 143, at 48; Garrett & Mitchell, *supra* note 43, at 923; PCAST Report, *supra* note 3, at 57–59; Max et al., *supra* note 136, at 283.

145. *See, e.g.*, NAS Report, *supra* note 4, at 25 ("A particular need exists for routine, mandatory proficiency testing that emulates a realistic, representative cross-section of casework, for example, DNA proficiency testing."); PCAST Report, *supra* note 3, at 59 ("[T]est-blind proficiency testing of forensic examiners should be vigorously pursued, with the expectation that it should be in wide use, at least in large laboratories, within the next five years."); Nat'l Comm'n on Forensic Sci., *supra* note 66, at 1 ("Proficiency testing is required of all accredited FSSPs. As a recognized quality control tool, it is the view of the Commission that proficiency testing should also be implemented by non-accredited FSSPs in disciplines where proficiency tests are available from external organizations.").

146. *See, e.g.*, Grzybowski *et al., supra* note 6, at 216–19; Grzybowski & Murdock, *supra* note 6, at 9; Lyn Haber & Ralph Haber, *Scientific Validation of Fingerprint Evidence Under Daubert*, 7 L., Prob. & Risk 87, 98 (2007); *United States v. Boone*, No. 23 Cr. 427 (PGG), 2024 WL 2160230, at *6–8 (S.D.N.Y. 2024).

reconsideration of one of the only judicial opinions to ever exclude fingerprint comparison evidence.[147]

Since the 1970s, such tests have primarily been supplied by CTS, which offers proficiency exercises annually across multiple forensic disciplines to thousands of examiners each year.[148] The sheer breadth of available data from these tests (more examiners and more guns tested) might meaningfully advance our understanding of the validity of firearms examination, even if they suffer from many of the same flaws as existing error-rate studies[149] as well as some additional deficiencies.[150] Unfortunately, however, signs abound (empirical and anecdotal) that existing proficiency tests are simply too easy to fulfill their various purposes. Members of law enforcement have admitted as much, with a former Scotland Yard fingerprint examiner even describing internal tests conducted by the FBI as "a joke," and testifying that "[i]f I gave my experts these tests . . . they'd fall about laughing."[151] Comments from examiners

147. *See* Simon A. Cole, *Grandfathering Evidence: Fingerprint Admissibility Rulings from Jennings to Llera Plaza & Back Again*, 41 Am. Crim. L. Rev. 1189, 1242–69 (2004) (discussing the dueling decisions in *United States v. Llera Plaza*, No. CR. 98–362–10, No. CR. 98–362–11, No. CR. 98–362–12, 2002 WL 27305 (E.D. Pa. Jan. 7, 2002) and *United States v. Llera Plaza*, 188 F. Supp. 2d 549 (E.D. Pa. 2002)).

148. *See* Collaborative Testing Servs., Inc., *About Collaborative Testing Services*, https://cts-forensics.com/index-forensics-testing.php [https://perma.cc/RDQ7-CFTJ](last visited May 22, 2026); Garrett & Mitchell, *supra* note 43, at 917; Max et al, , *supra* note 136, at 285; Jordan L. Peterson & Penelope N. Markham, *Crime Laboratory Proficiency Testing Results, 1978–1991, I: Identification & Classification of Physical Evidence*, 40 J. Forensic Sci. 994, 995–96 (1995).

149. Like those error-rate studies, CTS proficiency tests are declared rather than blind, anything but transparent with participant demographics, and have a volunteer and likely unrepresentative participant pool. *See, e.g.*, PCAST Report, *supra* note 3, at 57–59; Haber & Haber, *supra* note 146, at 95; Gardner et al., *supra* note 136, at 450; Koehler, *supra* note 84, at 1395–96; Kelley et al., *supra* note 140, at 120–21; Thompson, Black, Jain, & Kadane, *supra* note 143, at 151; Cole, *supra* note 136, at 1028–30; Ralph Haber & Lyn Haber, *Experimental Results of Fingerprint Comparison Validity & Reliability*: *A Review & Critical Analysis*, 54 Sci. & Just. 375, 386–87 (2014). The Houston Forensic Science Center is the exception, at least as regards conducting blinded proficiency tests. *See* Garrett & Mitchell, *supra* note 43, at 923; Robin Mejia et al., *Implementing Blind Proficiency Testing in Forensic Laboratories: Motivation, Obstacles, and Recommendations*, 24 Forensic Sci. Int'l Synergy 293 (2020); Maddisen Neuman et al., *Blind Testing in Firearms: Preliminary Results from a Blind Quality Control Program*, 29 J. Forensic Sci. 964 (2022).

150. Examiners often complete the tests collaboratively, complicating any effort to get at realistic assessments of *individual* competence. *See supra* note 136; Gardner et al., *supra* note 136, at 453 (3.1% of survey examiners worked on tests in groups, and 65.8% had their work checked by another examiner); Carroll et al., *supra* note 28, at 135 (11 of 15 surveyed examiners said they have proficiency test results verified); Stroman, *supra* note 118, at 158 (examiners may even receive hints and feedback from supervisors).

151. Cole, *supra* note 136, at 1249; *see also* Committee for the Advancement of the Science of Firearm & Toolmark Identification, *AFTE Response to the NACDL Task Force on the Future of Forensic Science*, 42 AFTE J 102, 103 (2010) (says of CTS tests: "could be more representative of casework compared to how most of them are currently prepared and administered"); Pauw-Vugts et al., *supra* note 32, at 117 ("Most of the participating examiners already participated in the CTS collaborative testing

participating in CTS testing routinely acknowledge the simplicity of comparisons: describing the tests as "quite easy," "[t]oo easy," "straightforward and very easy," and "a poor gauge of proficiency," while also going so far as to "suggest that you consider making the test more of a challenge in order to determine an error rate really reflective of actual casework in which borderline cases are not uncommon."[152] CTS itself has said essentially the same. In a statement opposing the use of their proficiency test results to estimate discipline-wide error rates, CTS described the comparison samples it uses as "more artifact than real world."[153] And recently, a representative for CTS outright conceded that the profit motive has resulted in the creation of easy and unrepresentative tests because their customers in the "forensic community prefe[r] that tests not be too challenging."[154] Scholars, unsurprisingly, have followed suit, "lambasting [proficiency tests] for involving less challenging samples than those encountered in casework"[155] and wondering "what these tests say about the proficiency of . . . examiners, if they say anything at all."[156]

program, but found this test insufficiently challenging to be of use in demonstrating competence in microscopy skills."); Stroman, *supra* note 118, at 158 (tests "typically do not accurately reflect the full range of difficulty experienced in real-life firearm and toolmark casework.").

152. Garrett & Mitchell, *supra* note 43, at 928; Adina Schwartz, *Challenging Firearms and Toolmark Identification—Part Two*, Champion 44 (2008); *see also* Kelley et al., *supra* note 140, at 122–23 (noting that of survey respondents who took a CTS latent print exam "participants found the questions to be relatively easy" & including one participant comment: "Latent prints were not challenging enough. The pattern was visible in every latent, which is not good. The latent prints from palms were very easy to locate too. In the future we would like to have more challenging test latents: latents where the pattern is not visible, only partial latents, etc."). Similar comments pervade additional CTS tests. *See, e.g.*, Collaborative Testing Servs., Inc., Firearms Examination Test No. 15–526, at 73 (2015) (comment saying "straightforward test"); Collaborative Testing Servs., Inc., Firearms Examination Test No. 14–527 57 (2014) ("Straight forward case"); Collaborative Testing Servs., Inc., Firearms Examination Test No. 14–526 81 (2014) ("The tests were very easy").

153. Collaborative Testing Servs. Inc., *CTS Statement on the use of Proficiency Testing Data for Error Rate Determinations* (2010) (on file with author). CTS makes similar caveats about the inadequacy of its tests to capture an error rate on each of its summary reports. *See, e.g.*, Collaborative Testing Servs., Inc., Firearms Examination Test No. 14527, at 1 ("the results compiled . . . are not intended to be an overview of the quality of work performed in the profession").

154. PCAST Report, *supra* note 3, at 57; Nat'l Comm'n on Forensic Sci., Views of the Commission Optimizing Human Performance in Crime Laboratories through Testing and Feedback 4 (2016); *Abruquah v. State*, 296 A.3d 961, 991–92(the same representative also said "that he has been under commercial pressure to make proficiency tests easier"), ("[A]lthough advocates of firearms identification tout periodic proficiency testing by Collaborative Testing Services Inc. . . . as a method of ensuring the quality of firearms identification, the record contains no evidence supporting efficacy of that testing. To the contrary, the evidence suggests that examiners rarely, if ever, fail CTS proficiency tests. Dr. Hamby confirmed that the industry's mandate to CTS with respect to proficiency tests 'was to try to make them [as] inexpensive as possible.'").

155. Max et al, , *supra* note 136, at 283.

156. Jason M. Tangen et al., *Identifying Fingerprint Expertise*, 22 Psych. Sci. 995, 995 (2011). Indeed, the list of publications that have criticized the simplicity of existing

But even beyond anecdotes, admissions, and the concerns of scholars, a growing empirical record (primarily developed in the realm of fingerprint comparisons)[157] also casts a substantial cloud of doubt over the utility of CTS proficiency tests. Research using software designed to assess the quality of latent prints has revealed that comparison samples in proficiency tests are overwhelmingly simpler and higher quality than the distorted and complex samples expected (and routinely encountered) in casework.[158] Bayesian Item Response

---

proficiency tests is staggering. *See* William A. Tobin & William C. Thompson, *Evaluating and Challenging Forensic Identification Evidence*, The Champion 19–20 (2006) (calling such tests "worthless as a true indicator of examiner proficiency); Haber & Haber, *supra* note 146, at 95 ("Further, the prints used in proficiency tests do not reflect normal casework. They are predominately or entirely of value in contrast to casework, in which the majority of latent prints are of no value."); Koehler, *supra* note 84, at 1395 ("CTS tests do not use realistic samples"); Thompson et al,, *supra* note 143, at 51 ("In their current form, proficiency tests have limited value for establishing the limits of reliability and accuracy that analytic methods can be expected to achieve as the conditions of forensic evidence vary. These tests provide little useful feedback to forensic analysts on the limits of their expertise when dealing with difficult cases or marginal evidence and hence have little value for helping experienced analysts hone and improve their skills."); Kelley et al., *supra* note 140, at 125 ("Primarily, findings suggest that current proficiency testing is not difficult"); Donald Kennedy & Richard Merrill, *Assessing Forensic Science*, 20 Issues in Sci. & Tech. 1 (2003) ("[P]ractitioners have not been subjected to rigorous proficiency testing, reliable error rates are not known"); Saks & Koehler, *supra* note 59, at 894 (describing proficiency testing as "infrequent, internal, and unrealistic"); Garrett & Mitchell, *supra* note 43, at 929 ("Our results reinforce the concerns of other scholars who have analyzed a subset of these CTS scores in the past and noted . . . simplicity of the test designs");Kelley et al., *supra* note 140, at 121 ("Clearly, there are indicators that current proficiency testing is perhaps too easy."); Jennifer L. Mnookin, *Of Black boxes, Instruments, and Experts: Testing the Validity of Forensic Science*, 5 Episteme 343 (2008) (opining that proficiency tests "have for the most part been extremely easy, far easier than the challenges that can be faced by examiners in actual casework.").

157. Work has also been conducted in the DNA sphere. *See* John M. Butler *et al.*, Nat'l Inst. of standards & Tech., NIST IR 8351, DNA Mixture Interpretation: A NIST Scientific Foundation Review,15–16, 97(2024) (reviewing existing proficiency tests for DNA over several years and concluding that "Current proficiency tests are primarily focused on single-source samples and simple two-person mixtures containing large quantities of DNA. To appropriately assess the ability of analysts to interpret complex DNA mixtures, tests of analysts should include the types of samples often seen in forensic casework, such as mixtures with low-template components and more than two contributors"). These findings have been echoed in litigation. *See United States v. Ferguson*, 2024 U.S. Dist. LEXIS 80118, at *20–21, 2024 WL 1928776 (D. Minn. 2018) ("Mr. Ferguson argues that the interpretations of BCA Lab scientists are unreliable, because the proficiency testing they are evaluated on uses less complex samples than those recovered from Items 4A and 4B. Dr. Hoogendoorn agreed that the proficiency tests used by the BCA 'are not very complex' and are not as complex as the mixtures involved in this case.").

158. *See* Kelley *et al.*, *supra* note 140, at 125 ("[T]he objective quality metric scores of the prints on the test suggest that the majority were of high ("good") quality and, thus, not especially challenging nor representative of latent prints that may be seen in casework"); Anthony J. Koertner & Henry J. Swofford, *Comparison of Latent Print Proficiency Tests with Latent Prints Obtained in Routine Casework Using Automated and Objective Quality Metrics*, 68 J. Forensic Id. 379, 379 (2018) ("the quality levels

Theory analysis of one test showed that its comparison samples were so easy they would not be able to so much as distinguish between professionals and novices.[159] And that statistical prediction bore out when three attorneys without professional training in fingerprint comparison took a CTS fingerprint proficiency test and completed it without committing any false positives.[160]

If the lawyers in the room find themselves frustrated, they should. The sole commentary received on the paper just discussed opposed its call for more rigorous proficiency testing by arguing that "if the bar exam was made so difficult so that recent law school graduates from accredited colleges and universities could not pass it, then only attorneys already in the bar could practice law."[161] But "bar exams . . . represent a comprehensive testing regime the likes of which [forensic] examiners need not navigate to join their profession . . . near-universal success . . . is not the norm."[162] Across multiple forensic disciplines, however, CTS simply has not risen to the same level of rigor. As Professors Brandon Garrett and Gregory Mitchell eloquently put it: "[j]ust as an easy exam (i.e., a test that all students can pass regardless of their levels of course knowledge and effort) is not a good measure of student expertise, an easy proficiency test can be a misleading measure of expert proficiency."[163] That examiners, near-universally, complete CTS tests without committing errors[164]

of latent fingerprints from proficiency tests are generally higher quality, less complex, and do not represent the quality levels observed in routine casework.").

159. *See* Luby & Kadane, *supra* note 136, at 120.

160. Indeed, my colleagues and I discovered along the way that the prints in the CTS test we took displayed abundant features for comparison (12–30 as opposed to the *de facto* threshold of eight that would generally allow an examiner to reach a conclusive identification or elimination), did not require us to navigate distorted or ambiguous prints, and included no close non-matches that might have provoked a misidentification. Max et al., *supra* note 136, at 286–92.

161. Gregory Eric Laskowski, *Letter to the Editor, Re: Assessing Latent Print Proficiency Tests: Lofty Aims, Straightforward Samples, and the Implications of Nonexpert Performance*, 69 J. Forensic Id. 397, 398 (2019).

162. Brendan Max et al, , *Author's Response, Letter to the Editor, Re: Assessing Latent Print Proficiency Tests: Lofty Aims, Straightforward Samples, and the Implications of Nonexpert Performance*, 69 J. Forensic Id. 400, 402 (2019); *see also* Nat'l Conf. of Bar Examiners, Ten-Year Summary of Bar Passage Rates, Overall and First-Time, 2015–2024, https://thebarexaminer.ncbex.org/2024-statistics/ten-year-summary-of-bar-passage-rates-overall-and-first-time-2015–2024 [https://perma.cc/WT3E-LP2E] (last visited July 17, 2025).

163. Garrett & Mitchell, *supra* note 43, at 922; *see also* John Butler, *NIST DNA Analysis Webinar Series: Validation Concepts and Resources:Part I Validation Overview*, Nat'l Inst. of Standards & Tech. (Aug. 6, 2014) *www.nist.gov/system/files/documents/forensics/01_ValidationWebinar-Butler-Aug2014.pdf* [https://perma.cc/TTQ5-XNUD]. (emphasizing the need for testing to "push the system until it fails in order to understand the potential limitations"),; 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) ("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."); Gutierrez, *supra* note 24, at 1556–61 (generally discussing the need for challenging testing and errors to reveal limits of validity).

164. *See, e.g.*, Garrett & Mitchell, *supra* note 43 at 927–28; Cole, *supra* note 136, at

thus means that such testing tells us little to nothing about discipline validity or individual examiner competence.[165] This situation, though, is not inevitable. Other examples of rigorous proficiency testing—testing that might truly inform decisionmakers—can be found beyond the law in the setting of clinical laboratories,[166] as well as (albeit isolated) in the forensic sphere.[167] Thankfully, we can now add CTS 23–5262 to that list.

II. Singularly Formidable & Fluky: CTS 23–5262's Predecessors, Samples, Participants, & Performance Measures

CTS 23–5262 clearly breaks from this long-established mold of meaninglessly simplistic testing. But it bears lingering on the ways that other CTS proficiency tests, specifically in the firearms examination realm, have stacked the deck in favor of practitioners and toward misleadingly low error rates to truly contextualize the significance of its deviation from the norm and the reason it should so impact judicial treatments of the field. Because performance on true eliminations has been the focus of courts[168]—as well as because a false positive (a misidentification) or a missed elimination poses the greatest threat of convicting the innocent[169]—this Article will focus exclusively on different source comparisons.[170] It calculates false positive rates either by including

1030–31; Jordan L. Peterson & Penelope N. Markham, *Crime Laboratory Proficiency Testing Results, 1978–1991, II: Resolving Questions of Common Origin*, 40 J. Forensic Sci. 1099 (1995); Cole, *supra* note 136, at 1212–15.

165. *See, e.g., United States v. Lewis*, 220 F. Supp. 2d 548, 554 (S.D. W. Va. 2002) (rejecting utility of proficiency tests that alleged experts "always passed").

166. *See* Garrett & Mitchell, *supra* note 43, at 915–16, 950–53 (describing proficiency testing regime under the Clinical Laboratory Improvement Act); Max *et al*, , *supra* note 136, at 293 (same).

167. *See* Pauw-Vugts et al. , *supra* note 32, at 117 (explaining that past errors "demonstrated a need for a competency-testing tool for firearms examiners. The test would need to be a proficiency test consisting of realistic and challenging comparisons"); John M. Butler, *NIST Interlaboratory Studies Involving DNA Mixtures (MIX05 and MIX13): Variation Observed and Lessons Learned*, 37 Forensic Science 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).

168. *See, e.g., United States v. Harris*, 502 F. Supp. 3d 28, 39 (D.D.C. 2020).

169. *See, e.g.*, PCAST Report, *supra* note 3, at 50; Cole & Sheck, *supra* note 91, at 819.

170. In terms of the presentation of the data, I largely follow the approach of meeting judicial and forensic audiences where they lay. *See* Gutierrez, *supra* note 24, at 1562–63. I sidestep concerns over confidence interval calculations (measurements of uncertainty that capture how low and high a measured error rate might truly be), *see* Cuellar *et al.*, *supra* note 30, at 9–11 & 17–19, as well as how to account for the outsized role / use of inconclusive responses during different-source comparisons, *see, e.g.*, Gutierrez, *supra* note 24, at 1551–53; Hofmann et al., *supra* note 81, at 323–25; Keith L. Monson et al., *Planning, Design and Logistics of a Decision Analysis Study: The FBI/Ames Study Involving Forensic Firearms Examiners*, 4 Forensic Int'l Synergy 100221, 3–4 (2022) (inconclusives in the denominator , i.e. treated as correct); PCAST Report, *supra* note 3, at 153 (remove inconclusives from the equation as a pass); Arkes & Koehler, *supra* note 92, at 161 (same); Dror & Scurich, *supra* note 87, at 334 (inconclusives in the numerator as well, i.e. treated as potentially incorrect)*;* Nicholas Scurich,

inconclusives in the denominator ("FPR AFTE"= false positives / false positives + inconclusives + eliminations) or omitting them entirely as a "pass" ("FPR PCAST" = false positives / false positives + eliminations).[171]

As it turns out, and in stark contrast to the realm of fingerprint comparisons, precious little work (some of it unreliable) has gone into analyzing performance on CTS proficiency tests for firearms examination. Scholars, as noted above, have collected comments from individual participants,[172] and issued general criticism of existing proficiency tests.[173] But previous summaries of participant performance leave much to be desired. The first, covering tests issued between 1978 and 1991, provides almost no information (including not breaking down whether a test involved bullets or cartridge cases) beyond the raw response data, from which we can calculate that participants achieved an AFTE FPR of 1.5%, a PCAST FPR of 1.8%, and specificity of 83.2%.[174] The second, from our old friend Mr. Murdock, gives even less: no information whatsoever (not even totals for each type of conclusion) beyond reported false positive rates (AFTE FPR style) of .9% for the years 1981–1997 and 1% for the years 1998–2002.[175] And the last, covering 1992–2005, reports an AFTE FPR of 1.5% and specificity of 57.1% overall (bullets and cartridge case tests) as well as an AFTE FPR of 2.5% and specificity of 43.4% for bullets specifically, but—beyond suffering the same data deficiencies described above—appears to contain glaring mathematical errors that make any reliance on its performance summary suspect to say the least.[176] Thus, despite nearly three decades of data, we learn little from these summaries other than that (like CTS tests writ large) they show rosy (and likely misleading) estimates of examiner

Inconclusives in Firearm Error Rate Studies Are Not 'A Pass', 21 L., Probab., & Risk 123, 125 (2022) (same); Alan Dorfman & Richard Valiant, *Inconclusives, Errors, and Error Rates in Forensic Firearms Analysis: Three Statistical Perspectives*," 5 For. Sci. Int'l: Synergy 100273, 6 (2022) (same). Instead, I use Clopper Pearson Exact 95% confidence intervals when necessary and provide false positive rates in terms proponents of firearms examination have accepted. *See* PCAST Report, *supra* note 3, at 153 (listing the Clopper Pearson method and providing a calculator); 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" 6 (2016) (discussing accepted error rate calculation methods).

171. *See* Gutierrez, *supra* note 24, at 1551–53,1562–63.
172. Schwartz, *supra* note 152.
173. *Id.*; Tobin & Thompson, *supra* note 156.
174. *See* Peterson & Markham, *supra* note 148, at 1018–19 (17 false positives, 175 inconclusives, 954 eliminations).
175. *See* Grzybowski et al., *supra* note 6, at 217.
176. Specifically, its total conclusions for bullet and cartridge case tests cannot be reconciled with the figures it provides combining the two. *See* Doug Murphy, *CTS Error Rates, 1992–2005 Firearms/Toolmarks*, 41st Ass'n. of Firearm & Tool Mark Examiners (AFTE) Training Seminar (Henderson, Nev., May 5, 2010) (total different source bullet comparisons, 2072, and cartridge case comparisons, 4851, do not sum to total overall comparisons, 9111), *available at* https://afte.org/wp-content/uploads/2024/11/swggun-cts-1992–2005.pdf.

performance. However comforting courts have found them,[177] they are tests in which near-universal passage (not to mention a dearth of *nuanced* data) teaches us little about discipline limits.

This Article extends and expands those summaries for tests (omitting briefly CTS 23–5262 as an outlier) between 2006 and 2024 covering bullet comparisons.[178] Across those nearly two decades, CTS released and reported the results of eighteen exercises involving bullets (seventeen of which involved elimination comparisons).[179] Admirably, the company covered multiple calibers,[180] and varied (as documented below in Table I) the make, model, and manufacturer of the guns it used to produce test samples (i.e., fired bullets).[181] But unfortunately, in keeping with the analysis in Part II.B. above, what it does not seem to have done is include challenging comparisons representative of the full range of casework difficulty. CTS does not explain (in Summary Reports or otherwise) precisely how or why it selects particular guns for its proficiency tests, but nowhere has it so much as alluded to making the types of conscious efforts (trolling databases, prescreening with algorithms, etc.) needed to include close nonmatches (to include the types of comparisons that would replicate the conditions under which a misidentification is likely to occur in the real world).[182] Indeed, though the method preferred by firearms examiners to capture difficulty (the inclusion of consecutively manufactured samples)[183] falls far short of the task,[184] CTS has done not even that across the last twenty years: consecutive serial numbers are an imperfect indicator of consecutive manufacture,[185] but even the precious few (four of eighteen) exercises that have forced

177. *See infra* Part III.A.

178. CTS makes only the last five years of summary reports available on its website. *See supra* note 134. A representative of the company, however, graciously provided me with summary reports dating back to 2006. All relevant summary reports remain on file with me. I will cite reports individually as becomes relevant throughout this section.

179. CTS's bullet comparison in 2020 involved only same-source comparisons. *See* Collaborative Testing Servs., Inc., Firearms Examination Test No. 20–5261 2 (2020), https://cts-forensics.com/reports/20–5261_Web-New.pdf.

180. *See e.g.*, Collaborative Testing Servs., Inc., Firearms Examination Test No. 06–527 2 (2006) (9mm Luger); Collaborative Testing Servs., Inc., Firearms Examination Test No.10–527 2 (2010) (40 caliber); Collaborative Testing Servs., Inc., Firearms Examination Test No. 12–527 2 (2012) (.38 caliber); Collaborative Testing Servs., Inc., Firearms Examination Test No. 13–527 2 (2013) (45 caliber).

181. Indeed, its use of Taurus, Smith & Wesson, & Ruger firearms mirrors the types of guns most recovered in relation to crimes in the United States. *See* Bureau of Alcohol, Tobacco, & Firearms, *Part III: Crime Guns Recovered & Traced Within the United States & its Territorie,* in National Firearms Commerce & Trafficking Assessment, Volume II: Crime Gun Intelligence & Analysis 18–22 (2023).

182. *See* Gutierrez, *supra* note 24, at 1559–61; Garrett, et al., *supra* note 36, at 145.

183. *See e.g.*, Best & Gardner, *supra* note 32, at 28; Monson et al., *supra* note 170, at 3.

184. *See generally* Gutierrez, *supra* note 24, at 1569–71.

185. *See* NRC Report, *supra* note 4, at 49 ("However, guns with consecutive serial numbers are generally not consecutively manufactured in full. Production of firearms is typically an assembly line process, drawing various preconstructed parts from large bins for assembly into a finished weapon. Hence, two firearms that bear consecutive serial numbers may have rolled off the line in sequence, but their frames, barrels, firing

examiners to confront elimination samples fired from the same make/model as known bullets have not utilized guns with consecutive serial numbers.[186]

Worse still, CTS has regularly stacked the deck in favor of accuracy in ways extreme enough to render its results meaningless or nearly so. Unlike determinations involving the assessment of individual characteristics, distinguishing bullets fired from guns with different class characteristics is allegedly "objective"[187] and should pose essentially no challenge to firearms examiners.[188] In the case of guns with different numbers of lands and grooves, it could be done by touch alone. This is why even critics of firearms examination have not objected to testimony limited to class characteristics,[189] and why study designers have generally avoided including out-of-class comparisons when evaluating error rates.[190] Yet despite the ways that including eliminations based on class characteristics alone would fail to provide a meaningful sense of practitioner accuracy, seven of the seventeen bullet exercises released by CTS involving eliminations over the last two decades (more than 41%) have revolved exclusively (or at least in part) around such comparisons.[191]

pins, and so forth need not have been manufactured right after each other.").

186. *See* Collaborative Testing Servs., Inc., Firearms Examination Test No. 07–526 2 (2007); Collaborative Testing Servs., Inc., Firearms Examination Test No. 13–536 2 (2013); Collaborative Testing Servs., Inc., Firearms Examination Test No. 18–526 2 (2018), Collaborative Testing Servs., Inc., Firearms Examination Test No. 19–527 2 (2019.

187. *Ricks v. Pauch*, No. 17–12784, 2020 U.S. Dist. LEXIS 50109, at *27 n.2 (E.D. Mich. Mar. 23, 2020).

188. *See* Gutierrez, *supra* note 24, at 1563.

189. *See Ross*, 129 N.Y.S.3d at 641; Scurich, Faigman, & Albright, *supra* note 92, at 5; NAS Report, *supra* note 4, at 8.

190. *See* Eric F. Law & Keith B. Morris, *Evaluating Firearm Examiner Conclusion Variability Using Cartridge Case Reproductions*, 66 J. Forensic Sci. 1704, 1706 (2021) (including class characteristic 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"). Empirical data backs up these concerns: examiners perform far better (although, disturbingly, not always perfectly) on class eliminations. *See* Best & Gardner, *supra* note 32, at 32–35 (examiners achieved specificity of 84.1% on class eliminations, missing 7 of 44, but only 13.1% on within-class eliminations, missing 153 of 176); Keisler et al., *supra* note 97, at 56, 58 (examiners achieved 100% specificity on out of class comparisons versus only 73.1% specificity on within-class comparisons). But there are exceptions. *See* Gil Hocherman & Pavel Giverts, *Identification of Polygonal Barrel Sub-Family Characteristics*, 35 AFTE J. 197, 200 (2003). And scholars have begun to raise concerns about whether examiners indeed have consensus definitions of what constitutes a class characteristic difference. *See* Nicholas Scurich & Hal Stern, *Commentary on: Monson KL, Smith ED, Peters EM. Accuracy of Comparison Decisions by Forensic Firearms Examiners*, 68 J. Forensic Sci. 1093 (2023).

191. *See* Collaborative Testing Servs., Inc., Firearms Examination Test No. 11–526 16 (2011) (class differences of six lands and grooves versus eight lands and grooves); Collaborative Testing Servs., Inc., Firearms Examination Test No. 12–527 10 (2012) (class differences of six lands and grooves versus five lands and grooves); Collaborative Testing Servs., Inc., Firearms Examination Test No. 14–526 14 (2014) (one elimination displayed class difference in terms of land impression widths); Collaborative Testing Servs., Inc., Firearms Examination Test No. 16–526 13 (2016) (class differences of six lands and grooves versus nine lands and grooves);

Table I: Firearm Makes & Models Used in CTS Bullet Comparison Proficiency Tests (Exempting 23–5262) from 2006–2024

Test	Knowns	Eliminations
06–527	Ruger P89	Ruger P85
07–526	Colt Trooper MKIII 357 Magnum CTG	Colt Trooper MKIII 357 Magnum CTG
08–527	Llama Large Frame	Llama Model IX-A
09–526	Ruger P85 MKII	Ruger P89 DC
10–527	Springfield XD-40	Springfield XD Compact
11–526	Beretta 92FS; Beretta 92D	Hi-Point 995
12–527	Kel-Tec P3AT	Walther PPK/S-1
13–527	Ruger P90	Ruger P90
14–526	Ruger P95DC	Taurus PT 24/7; Ruger P85 MK II
15–527	Ruger P94DC	Ruger P91DC
16–526	Colt MK IV Series 80 Mustang	Hi-Point CF
17–527	Ruger SR9C	Ruger SR9
18–526	Beretta PX4 Storm	Beretta PX4 Storm
19–527	SCCY CPX-2	SCCY CPX-2; Smith & Wesson MP
20–5261	Smith & Wesson M&P Bodyguard	N/A
21–5262	Springfield XD-40	Taurus Millennium PT140 Pro
22–5261	Beretta 92FS	Beretta 92F; Springfield XD-9 4.0 Mod
24–5261	Tanfoglio Witness	Hi-Point 995

As documented below in Table II, all this has predictably led to the type of universal passage that teaches labs and courts so little about examiner competence and discipline limits.[192] Though CTS tests hundreds of partic-

Collaborative Testing Servs., Inc., Firearms Examination Test No. 19–527 16–17 (2019) (class differences for some eliminations between five lands and grooves versus seven lands and grooves); Collaborative Testing Servs., Inc., Firearms Examination Test No. 22–5261 20 (2022) (one elimination displayed class difference in terms of land impression widths); Collaborative Testing Servs., Inc., Firearms Examination Test No. 24–5261 14 (2024) (class differences of six lands and grooves versus eight lands and grooves).

192. One relevant figure that Table II does not report (because CTS included this information in Summary Reports only sporadically) is attrition (also called the non-response rate) which measures the percentage of participants who ordered a test but did not return responses. Failing to account for non-response / missing data can result in misleadingly rosy estimates of error unreflective or the true performance and accuracy of practitioners in the real world. *See* Khan & Carriquiry, *supra* note 105, at 4–5. If high enough (over 20% in the opinion of many scholars) attrition can outright invalidate results. *See generally* Jo C Dumville et al., *Reporting Attrition in Randomised Controlled Trials*, 332 BMJ 969 (2006). CTS reported attrition in each Summary Report between 2006 and 2013, during which it ranged from as low as 8% to as high as 23%. *See* Collaborative Testing Servs., Inc., Firearms Examination Test No. 13–526 1 (2013); Collaborative Testing Servs., Inc., Firearms Examination Test No. 8–527 1 (2008). CTS did not thereafter report this figure in Summary Reports, but it did supply a retrospective look at attrition from 2018–2023 when it analyzed the

ipants each year, it rarely identifies poor performers (i.e., participants who commit false positives): never (exempting CTS 23–5262) has it failed more than eleven participants (3.7%), and on more than half its exercises less than 1% bombed.[193] Examiners have even achieved specificity levels (as high as 100% and above 80% on fourteen of seventeen tests involving eliminations) they could brag about. Tests so easy that essentially everyone passes cannot distinguish between examiner competence levels, help labs improve policies, or educate courts readying themselves for admissibility decisions.

Table II: Error & Accuracy Rates for CTS Bullet Elimination Comparisons 2006–2024 (Exempting 23–5262)

Test	Within-Class	%/# Poor Performers	FPR AFTE	FPR PCAST	Specificity	Inconclusive Rate
06–527	Yes	2.6% / 4	2.4%	3.7%	61.4%	36.2%
07–526	Yes	.7% / 2	.9%	1.2%	71.6%	27.6%
08–527	Yes	1.8% / 3	1.2%	1.4%	81.4%	17.4%
09–526	Yes	0% / 0	0%	0%	64.9%	35.1%
10–527	Yes	2.5% / 5	2.3%	2.7%	83.3%	14.4%
11–526	No	0% / 0	0%	0%	100%	0%
12–527	No	0% / 0	0%	0%	100%	0%
13–527	Yes	3.7% / 11	3.3%	4%	80.6%	16.1%
14–526	Some	1.8% / 7	1.2%	1.5%	80.2%	18.6%
15–527	Yes	.7% / 2	.7%	.9%	84%	15.3%
16–526	No	0% / 0	0%	0%	99.7%	.3%
17–527	Yes	.7% / 2	.7%	.9%	80.4%	18.9%
18–526	Yes	0% / 0	0%	0%	92.4%	7.6%
19–527	Some	2.1% / 5	1.4%	1.6%	88.8%	9.6%
20–5261	-	-	-	-	-	-
21–5262	Yes	2.6% / 7	2.6%	3%	85.5%	11.8%
22–5261	Some	.6% / 2	.4%	.4%	92%	7.6%
24–5261	No	.2% / 1	.1%	.1%	99.9%	0%
Totaled All	-	-	.9%	*1.1%*	86.1%	13%
Totaled Out-Class	-	-	.1%	.1%	98.9%	1%
Totaled In-Class	-	-	1.2%	1.5%	81.2%	17.6%

results of CTS 23–5262; attrition during that period ranged from 11% - 23%, with the highest rate of 24% coming on CTS 23–5262 itself. *See* CTS 23–5262, *supra* note 23, at frequently asked questions 4.

193. Even these errors—though it would be inappropriate to exclude / forgive them after the fact on this basis, *see* PCAST Report, *supra* note 3, at 95—may be more clerical than substantive. *See* Murphy, *supra* note 176 (identifying potential causes of error as including "Mislabeling of evidence by examiner," "Mislabeling of evidence by CTS," and "Mistake in report or notes").

There is yet one more way, however, to capture the underhanded-pitch-simplicity of CTS's bullet comparison proficiency tests: we can compare them to more formal validation and error-rate studies in the firearms examination field. There have been three such studies to date involving bullets that have produced error rate estimates (at least point estimates) in line with claims of firearms examiners (i.e., around or less than 2%).[194] Those studies, as I have argued at length elsewhere and address in further detail in Part III.B., likely provide gross underestimates of the true rates of error and accuracy in casework largely because they failed to make any conscious and sufficient efforts to seek out close nonmatch elimination samples.[195] Yet they managed to pose far more of a challenge to their participants than CTS tests historically ever did.

To demonstrate, Figures I and II compare the point estimates and Clopper Pearson Exact 95% confidence interval ranges for the aggregated results of the three error rate studies ("Blackbox Studies")[196] and the aggregated results of CTS bullet comparison tests from 2006–2024 (again exempting CTS 23–5262). Only considering AFTE FPRs are the results of the Blackbox studies and CTS tests essentially indistinguishable (i.e., have confidence interval ranges that substantially overlap),[197] otherwise CTS participants massively outperformed those of the Blackbox studies (indeed the inconclusive rate and specificity figures are not even close). It is, of course, impossible to say precisely why this occurred. Maybe, for example, the participants in CTS were just far better than the examiners who participated in the Blackbox studies,[198] or perhaps collaboration and verification during CTS testing proved out the old adage that two heads are better than one.[199] But one possible explanation (a more likely one,

194. *See* Best & Gardner, *supra* note 32, at 35; Hicklin et al., *supra* note 113, at 6; Keith L. Monson et al., *Accuracy of Comparison Decisions by Forensic Firearms Examiners*, 68 J. Forensic Sci. 86, 86 (2023).

195. Gutierrez, *supra* note 24, at 1561–73; *see also supra* notes 106–111.

196. This label mirrors that used elsewhere to concisely name such conscious efforts to estimate error rates. *See e.g.*, PCAST Report, *supra* note 3, at 506; Khan & Carriquiry, *supra* note 105, at 1.

197. *Compare* Hicklin et al., *supra* note 113, at 13 (using confidence interval overlap to gauge differences, or lack thereof in performance); Gutierrez & Prokesch, *supra* note 13, at 1341–42 (same). I use this method to avoid issues that have plagued the literature surrounding p-values and arbitrary thresholds for statistical significance. *See generally* Martin J. Gardner & Douglas G. Altman, *Confidence Intervals Rather Than P Values: Estimation Rather Than Hypothesis Testing*, 292 BMJ 746 91986); Ronald L. Wasserstein RL & Nicole A. Lazar, *The ASA Statement on P-Values: Context, Process, and Purpose*, 70 Am. Stat. 129 (2016).

198. This explanation is unlikely. In one study, 94% of participants had taken a proficiency test within the last year (making them likely participants in CTS testing as well. *See* Hicklin et al., *supra* note 113, at 3. And in another, no less than 88% (and more likely 98–100%) of participants hailed from accredited labs, *see* Monson, Smith, & Peters, *supra* note 194, at 88; Gutierrez & Addyman, *supra* note 31, at 1100, meaning they too took proficiency tests and were therefore likely CTS participants, *see supra* note 141.

199. This explanation is also unlikely. Although some labs allow for collaboration and/or verification during proficiency testing, *supra* note 150, so too did one of the studies, *see* Best & Gardner, *supra* note 32, at 34–35 (explaining that "[t]o further replicate real casework, participants were asked to follow their laboratory's policies and procedures

given the issues with CTS testing generally and specifically in the realm of firearms examination) is that CTS simply used far easier comparison samples, like bullets that posed little risk of duping examiners into a misidentification, or bullets rich enough in information and different enough from one another that CTS participants achieved accuracy levels (high specificity and `low inconclusive rates) on elimination comparisons unheard of elsewhere in the testing of practitioners.[200] Though examiners have continued to cite these results in defense of their chosen discipline,[201] there is therefore no legitimate justification for anyone else to follow suit.

Figure I: Point Estimate & 95% Clopper Pearson Exact Confidence Interval Comparisons of False Positive Rates Between Cumulative CTS Results & Combined Validation Study Results

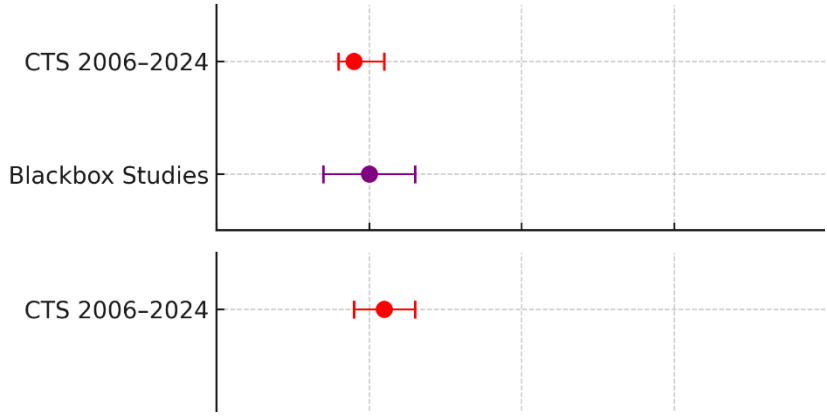
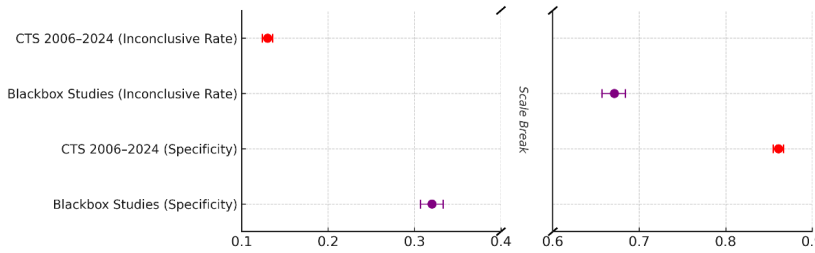


Figure II: Point Estimate & 95% Clopper Pearson Exact Confidence Interval Comparisons of Specificity & Inconclusive Rates Between CTS Results & Combined Validation Studies



It does not appear that CTS set out to alter this course when it designed and distributed its 2023 bullet comparison proficiency test. The company did

to include any quality control systems and verification" meaning that at least 68% of participants had their work checked by a second examiner).

200. *See, e.g.*, Sinha & Gutierrez, *supra* note 78, at 133–34 (collecting inconclusive rates across multiple firearms examination error rate studies).

201. *See, e.g.*, Todd J. Weller, *Declaration of Todd J. Weller on behalf of the Montgomery County District Attorney's Office* 24–25 (June 11, 2021) (on file with author).

nothing to remedy the myriad limitations, discussed above, about the declared nature of its tests, its transparency about participants, and the potential for collaboration.[202] Indeed, when releasing its Summary Report for the exercise, the company not only exhaustively documented the ways it followed its normal test design procedures, but also admitted outright that the "test was more challenging than originally intended."[203] Even that caveat, however, undersells the drastic gulf separating CTS 23–5262 from the previous fifty-plus years of bullet proficiency tests. The test used a CZ 75 P-0740 caliber firearm to create its known bullets and included four elimination comparisons: three of them to a CZ 40B firearm (within-class) and one to a Desert Eagle 40 S&W (out-of-class).[204] Excluding the latter as uninformative (examiners should not have struggled to differentiate the cut rifling of the CZ firearms from the polygonal rifling of the Desert Eagle),[205] performance, as summarized below in Table III, looked nothing like the results of past tests.[206]

Table III: Error & Accuracy Rates for CTS 23–5262 Within-Class Elimination Comparisons

Participant Group	% / # Participants	AFTE FPR	PCAST FPR	Specificity	Inconclusive Rate
All	100% / 280	19.3%	37.4%	32.3%	48.5%
Poor Performers	21.9% / 61	88.5%	94.2%	5.5%	6%
PP Subgroup 1	17.1% / 48	100%	100%	0%	0%
PP Subgroup 2	1.9% / 5	66.7%	71.4%	26.7%	6.7%
PP Subgroup 3	2.9% / 8	33.3%	57.1%	25%	41.7%

Rather than identifying almost no poor performers, CTS 23–5262 provoked misidentifications by sixty-one people (a whopping 21.9% of participants and close to the percentage of participants who correctly eliminated on all comparisons, eighty-seven participants or 31%); most misidentified on all three within-class comparisons (PP subgroup 1) while smaller numbers

202. CTS 23–5262, *supra* note 23, at 3 & frequently asked questions.
203. *Id.*
204. *Id.* at 2–3; Carroll et al., *supra* note 28, at 129.
205. *Supra* note 190; *see also* NRC Report, *supra* note 4, at 32 & 46 (describing stark differences in the sharp versus smooth edges of cut and polygonal rifled barrels). Indeed, participants almost entirely did not struggle with the out of class comparison. Adjusting for participant WD8HVT, *see infra* note 231, no examiner reached a false positive conclusion, 275 examiners eliminated, and only three examiners were inconclusive. *See* CTS 23–5262, *supra* note 23, at 10. But one additional examiner (9GC4WM), while inputting no conclusions for that comparison, did indicate in their comments that they were eliminating. *Id.* at 20–21. This makes the specificity for the out of class elimination near perfect (98.9%) and the inconclusive rate vanishingly low (1.1%). Repeatedly, examiners were able to reach those eliminations by identifying the difference between cut and polygonal rifling. *See e.g., id.* at 11, 14, 16, & 24.
206. Table I utilizes the response summary provided by CTS. *Id.* at 10. But it adjusts those figures because one participant, WD8HVT, committed an obvious (and undeniable) scrivener's error. Based on their notes they clearly intended to eliminate on the out of class comparison and to identify item 5 as having been fired by the same gun as the knowns. *Id.* at 52.

misidentified on either two (PP Subgroup 2) or one of those comparisons (PP subgroup 3). Instead of serving up the low false positive rates law enforcement has so fervently touted,[207] CTS 23–5262 resulted in an AFTE FPR of nearly 20% and a PCAST FPR approaching 40%; indeed, even the most favorable misidentification rate from the exercise (the Clopper Pearson Exact 95% lower bound of a confidence interval for the AFTE FPR) was still 16.7%, and the misidentification rate might have been as high (the Clopper Pearson Exact 95% upper bound of a confidence interval for the PCAST FPR) as 42.2%. And on the accuracy side of the coin, not only did specificity plummet from unprecedented highs (for firearms examination anyway) above 80% to a rate below even that measured in blackbox studies, but the inconclusive rate skyrocketed from the low double digits (unheard of in said blackboxes) to nearly 50%.[208]

Worse still, those poor performers who avoided misidentifying across the board, in many cases, did so quite narrowly: four of the six noted in their comments that they observed similarity between the knowns and the elimination comparisons (i.e., a lean towards a misidentification or an Inconclusive A opinion)[209] but simply did not consider it sufficient for a more definitive conclusion.[210] If we account for test-taking bias—and assume that such comparisons might have been pushed to misidentifications with casework incentives at play[211]—that best case scenario AFTE FPR lower bound becomes 17.5% and that worst case scenario PCAST FPR upper bound becomes 43.8%. And finally, perhaps most damning of all, not even the participants we would expect to perform best (those working professionally in accredited labs) managed to navigate the test with error and accuracy rates mirroring the claims to near-certainty emanating from the discipline's defenders. True, as firearms examiners reviewing the results of CTS 23–5262 were quick to point out, that group performed substantially better than the remainder of participants on the test (larger percentages of them eliminated correctly and smaller percentages of them were duped into a misidentification).[212] But their misidentification rates were nevertheless far from comforting, exceeding what some have set as the limit of method

207. *Infra* note 241.
208. *Compare* Monson, Smith, & Peters, *supra* note 194, at 89 (reporting specificity for bullet comparisons of 33.8% and inconclusive rate of 65.55).
209. *See supra* note 77; Gutierrez & Addyman, *supra* note 31, at 1097–98 (describing the range of potential inconclusive conclusions in more detail and explaining the need to investigate leans towards false identifications).
210. CTS 23–5262, *supra* note 23, at 32–33, 62, 69, & 72 (those participants described "an agreement of individual characteristics, but insufficient for an identification," "some similar individual characteristics but not enough to identify," and twice "a considerable degree of similarity").
211. *See generally* Scurich et al., *supra* note 92.
212. Carroll et al., *supra* note 28, at 131–32.

validity.[213] They may have achieved a false positive rate no better than 3.1% and possibly as bad as 25.6%.[214]

But how—if neither CTS nor firearms examiners invested in minimizing the exercise's implications could identify changes in participants, test-taking procedures, or conscious test-design choices that would explain such deep and wide-ranging performance failures[215]—did this come to pass? It appears that the test's comparison samples (through a failure of quality control) simply included challenges both fair and representative of casework, but also (disturbingly) insurmountable for most examiners. Per its normal test design procedures, CTS outsourced production of comparison samples for CTS 23–5262 to an accredited laboratory. The production included quality measures such as random checks of 10% of created samples for "consistency, reproducibility, and quality," as well as pretesting by five qualified examiners.[216] And by those measures, nothing seemed seriously awry: Samples were rejected at a "not atypical" rate of 1% for the knowns and a slightly higher 3% for the within-class eliminations. Though pretesting examiners noted some sources of difficulty—with one calling it "the most difficult PT they ha[d] ever received" and others pointing out the within-class nature of the eliminations, the subtlety of available markings (only "finer-striated" marks), and the minimal "reproducibility" of those markings even as between the known bullets—they also described the samples as displaying "sufficient markings for conclusions," and called the test itself both "practical" and "more similar to casework."[217] Yet despite these checks, participants complained time and again about two major issues: (1) the bullets they were provided (including knowns) displayed poor "quality and reproducibility

213. *See* PCAST Report, *supra* note 3, at 151–52 ("To be considered reliable, the FPR should certainly be less than 5 percent and it may be appropriate that it be considerably lower, depending on the intended application"). Though not specified, that figure likely applies to the upper bound for a false positive rate. *See id.* at 153 ("Upper 95 percent one-sided confidence bounds should thus be used for assessing the error rates and the associated quantities that characterize forensic feature matching methods.").

214. CTS provided firearms examiners reviewing the test not with the conclusion totals for participants from accredited labs, but instead merely with the number of such participants who eliminated on all within class comparisons, reached inconclusive for all within-class comparisons, or misidentified at least one within class comparison. Carroll et al., *supra* note 28, at 141–43. I requested that data (even in anonymized form) via email, but a CTS representative refused to release it. That creates complications because the sixteen examiners from accredited labs who committed false positives might have belonged (at worst) entirely to PP Subgroup 1, and (at best) might have been split between all three PP Subgroups (i.e. accounting for all of PP Subgroups 2 and 3 with three left over from PP Subgroup 1). To calculate a worst-case scenario, I assumed the former and provide above the Clopper Pearson Exact 95% upper bound for the PCAST FPR. To calculate a best-case scenario, I assumed the latter and provide above the Clopper Pearson Exact 95% lower bound for the AFTE FPR. So there is no doubt about such performance violating validity thresholds, *see* PCAST Report, *supra* note 3, at 151–52, readers should know that the upper bound for the best-case scenario is a false positive rate of 6.8%.

215. *See* CTS 23–5262, *supra* note 23, at frequently asked questions; Carroll et al., *supra* note 28, at 141–43.

216. *See* CTS 23–5262, *supra* note 23, at frequently asked questions 1–3.

217. *Id.*

of marks/striae,"[218] i.e., examiners observed "few individual characteristics";[219] and (2) many of the bullets showed damage from having struck the bottom of the water tank into which guns are fired to (in theory, but not, in this case, in practice) ensure pristine samples.[220]

None of this means, however, that CTS 23–5262 set an unfair, unrealistic, or uninformative challenge for participants. Most bullets examiners encounter in casework display damage of some kind,[221] and it is similarly commonplace for examiners to grapple with guns that mark poorly or inconsistently.[222]

---

218. CTS 23–5262, *supra* note 23, at 2–3.

219. Carroll et al., *supra* note 28, at 133.

220. Water tanks are meant to ensure pristine bullets, and yet damage can still occur. *See* NRC Report, *supra* note 4, at 76 (In forensic laboratories, firearms can be discharged into water tanks . . . so that bullets may be recovered with no damage other than those left by the firing process); James E. Hamby et al., *The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries*, 41 AFTE J. 99, 110 (2009) (defining "tank rash" as "an unofficial term used by some firearms examiners to denote the damage caused when fired bullets strike the bottom of the water recovery tank"). Both CTS and firearms examiners reviewing CTS 23–5262 describe multiple comments from participants attesting that bullets in the exercise displayed this type of damage. *See* CTS 23–5262, *supra* note 23, at 3 ("[c]omments concerned . . . tank damage noted on both known and questioned bullets"); Carroll et al., *supra* note 28, at 133 ("[g]eneral comments from participants included . . . [t]he condition of the bullets was poor in some instances, to include abrasions on the ogive and bearing surface, suspected to be water tank damage.").

221. *See e.g.*, Best & Gardner, *supra* note 32, at 31 ("Once bullets are fired . . . they may incur damage as a result of striking a target. This is common in casework and as such contributes to the difficulty of bullet comparisons. Damage to the bullet can result in foreshortening and partial or complete obliteration of the individual characteristics; however, identification is often still possible with damaged bullets depending on what area and to what extent the bullet is damaged. It is important to conduct validation studies using specimens for comparison that are representative of casework"); NRC Report, *supra* note 4, at 76 ("A basic, fundamental challenge of firearms identification is that crime scenes are not necessarily . . . controlled settings. The task of pattern matching between pieces of bullet or cartridge case evidence . . . is already a difficult one. But it can be made more difficult by the nonpristine nature of crime scene evidence. Bullets can strike wood, asphalt, human tissue and bone, or other substances and, consequently, can be seriously warped, deformed, or fragmented; they can be lodged such that even their recovery for analysis can cause damage . . . The possible differences between pristine exhibits and crime-scene samples are a continuing challenge for conventional firearms identification"); Thompson, *supra* note 69, at 26 (noting that considerations of damage are "especially important in the evaluation of bullets"); Zhe Chen et al., *Pilot Study on Deformed Bullet Correlation*, 306 Forensic Sci. Int'l 110098, 1 (2020) ("bullets found at a crime scene may be deformed or fragmented due to impact, resulting in distorted, expanded, or missing striated toolmarks, which can cause challenges for comparisons performed by firearm examiners or algorithms").

222. *See e.g.*, NRC Report, *supra* note 4, at 193 (noting difference in the extent to which barrels leave useful markings on fired bullets, with many weapons "so finely polished and machined as to render toolmarks too subtle to readily distinguish"); Weller, *supra* note 201, at 11–12 (collecting research to defend the hesitance of firearms examiners to render elimination conclusions given that "some know-matching samples will be poorly-marked"). Indeed, statistical studies of markings on fired bullets have revealed

Indeed, not only did the test fail to trip up one automated comparison algorithm fed its comparison samples, but examiners reviewing the test for AFTE even believed that CTS 23–5262 should serve as a model for labs to follow to create training sets with "a low quality of striated marks."[223] Instead, promulgating a realistically challenging proficiency test well served needs internal to labs and the discipline: helping course-correct misunderstandings about the use of inconclusive as a response, "reduc[ing] the chance of similar occurrences [misidentifications] in the future, resulting in better trained examiners who are less likely to err," corroborating concerns about examiners being "influenced by factors outside their analysis of the evidence,"[224] and even identifying the troubling tendency of some examiners to (contrary to industry guidance)[225] "conflat[e] lack of disagreement with significant/sufficient agreement."[226] By producing a test better representative of the full range of casework difficulty,[227] CTS 23–5262 will also (albeit in ways less salutary to the field of firearms examination) serve litigants and the courts when in search of meaningful error rates to inform method admissibility and examiner qualification inquiries.

III. HAVING HEARD THE BELL RUNG: THE PATH FORWARD FOR LITIGANTS & THE COURTS

Federal Rule of Evidence 702 sets the framework for how courts and litigants should react to CTS 23–5262. It requires judges to assess the qualifications of expert witnesses, the reliability[228] of the "principles and methods"

---

that poor marking can mean that the level of agreement in known non-matching bullets exceeds that of known matches. *See* Alfred A. Biasotti, *A Statistical Study of the Individual Characteristics of Fired Bullets*, 4 J. FORENSIC SCI. 34, 38–40 (1959); 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).

223. *See* Carroll et al., *supra* note 28, at 138 & 142. Participants in the test expressed more displeasure, frequently describing it as too difficult. CTS 23–5262, *supra* note 23, at 62, 75–76, & 81. At least one, however, disagreed. *Id.* at 77 ("The test was not hard – it was very easy to render a conclusion of inconclusive – not because I couldn't tell, but because it was the right answer").
224. *See* Carroll et al., *supra* note 28, at 141–42.
225. *See* Comm. for the Advancement of the Sci. of Firearm & Toolmark Identification, *supra* note 7, at 287.
226. *See* Carroll et al., *supra* note 28, at 141–42.
227. *Supra* notes 92 & 106.
228. Rule 702's use of "reliable" does not quite align with the way that scientists deploy that term, and the Supreme Court did not fully resolve this conflict by clarifying that "*evidentiary reliability* will be based on scientific validity." *Daubert*, 509 U.S. at 590 & 591 (emphasis added). As touched on above, to a scientist or mathematician, reliability is (usually) synonymous with consistency: the rate at which the same examiner reaches the same conclusion, when looking at the same evidence on different occasions (repeatability) and the rate at which different examiners looking at the same evidence reach the same conclusion (reproducibility). OSAC HUMAN FACTORS REPORT, *supra* note 88, at 4. Scientific validity requires both consistency (reliability) and accuracy (how often an examiner of method reaches a conclusion that aligns with ground truth). *Id.* at 3–5 ("validity of . . . a method can be assessed by testing whether it consistently produces accurate results when applied to test specimens of known

underlying their testimony, and whether (in a given case) "the expert's opinion reflects a reliable application" of such principles and methods[229] (what some scientific scholars have characterized as "foundational validity" and "validity as applied").[230] Judges, especially in criminal cases, have not always fulfilled this gatekeeping role impartially or with gusto.[231] But recent amendments to the rule—amendments spurred on in large part by judicial treatment of forensic evidence—have emphasized the *proponent's* burden to demonstrate the legitimacy of the expert testimony they hope to supply, as well as the inadequacy of judicial willingness to simply dispense with challenges to questionable techniques by cabining all criticisms to issues of weight as opposed to admissibility.[232] As such, there is reason to hope that the trend in courts towards more skeptical and searching review of firearms examination[233] evidence will not only continue but expand. To that end, the extreme and nigh-unprecedented false positive rate from CTS 23–5262 (as well as its glaringly low specificity rate) provides litigants and the courts with data critical to each of Rule 702's threshold admissibility criteria. Courts comforted by claims regarding the *de minimis* probability of misidentifications by firearms examiners when considering "known or potential rate of error" under *Daubert*[234] will need to reassess

source or type"). This article will continue to generally use the terms validity and reliability in their scientific sense. *See* Nat'l Inst. Standards & Tech., *supra* note 53, at 9.

229. Admissibility standards vary across the fifty states, *see e.g.*, Stofli, *supra* note 27, at 862, but those different tests do not appear to much impact ultimate admissibility decisions given that an inquiry into the general acceptance of a method in the relevant scientific community will so often dictate what judges think about the remainder of the factors in *Daubert*. *See* Adina Schwartz, *A Dogma of Empiricism Revisited*, 10 Harv. J. Law & Tec 149, fn.176 (1997); Jennifer L. McGarrity, *Facilitated Communication-Just Another Admissibility of Evidence Issue for Courts?*, 20 Dayton L. Rev. 935, 946 (1995).

230. PCAST Report, *supra* note 3, at 43 ("we have adopted specific terms to refer to the *scientific* standards for two key types of scientific validity, which we mean to correspond, as scientific standards, to the legal standards in Rule 702 . . . by 'foundational validity,' we mean the *scientific* standard corresponding to the legal standard of evidence being based on 'reliable principles and methods,' and . . . by 'validity as applied,' we mean the *scientific* standard corresponding to the legal standard of an expert having 'reliably applied the principles and methods'") (emphasis original).

231. *See, e.g.*, Erin Murphy, *Neuroscience and the Civil/Criminal Daubert Divide*, 85 Fordham L. Rev. 619, 621–624 (2016); Peter J. Neufeld, *The Near Irrelevance of Daubert to Criminal Justice and Some Suggestions for Reform*, 95 S1 A. 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); 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).

232. Fed. R. Evid. 702, Advisory Committee Comments (2023); John D. Bates, Memorandum from the Honorable John D. Bates, the Chair of the Committee on Rules of Practice and Procedure, to Scott S. Haris, the Clerk of the Supreme Court 227 (Oct. 19, 2022).

233. *Supra* notes 36–38.

234. *Daubert*, 509 U.S. at 593–94 (also emphasizing other factors: a method's testability,

their untroubled reliance on accuracy rates proffered by law enforcement and practitioners. Courts will need to take seriously issues raised by litigants about whether firearms examination methods rest now or can ever find solid ground when applied to damaged bullets. And, in an unknown number of cases involving the dozens of participants who failed CTS 23–5262, calls for more rigorous and empirical assessments of the qualifications of forensic "experts"[235] should find new purchase.[236]

A. *Reassessing Assumptions About False Positive Rates*

First and perhaps most obviously, CTS 23–5262's results should unseat the judiciary's current tendency towards uncritical acceptance of industry claims regarding the false positive rate associated with firearms examination.[237] The U.S. Supreme Court in *Daubert* specifically directed courts to consider "the known or potential rate of error" when evaluating the reliability of challenged methodologies under Rule 702.[238] Research scientists and statisticians have similarly emphasized the centrality of error rates (especially false positive rates) to assessments of method performance,[239] and even forensic practitioners have largely fallen in line with these views.[240] Unsurprisingly, then, debates over how to appropriately estimate the false positive rate for bullet and cartridge case comparisons, as well as what figure to ultimately rely on, have featured prominently in the wave of admissibility challenges to firearms examination that has confronted the courts in recent years.[241] In these debates, firearms examiners and law enforcement groups have consistently claimed, through *ipse dixit* or by referencing proficiency testing results or those of validation studies, that trained analysts will commit misidentifications, at worst, 2% of the time they confront ammunition from different sources.[242] But, as discussed above, scientists and

peer review, standards / controls, and general acceptance).

235. *See, e.g.*, Garrett & Mitchell, *supra* note 43, at 910; Edward J. Imwinkelreid, *Revisiting "The Proficiency of Experts": The Need to identify & Analyze the Evidentiary Sub-Issues*, 42 Rev. Litig. 1, 6 (2022) (calling The proposal by Professors Garrett and Mitchell "that the courts mandate proof of the witness's proficiency testing . . . as straightforward as it is sensible").

236. Imwinkelreid, *supra* note 235, at 7–8 (suggesting that proficiency testing data would be relevant across each of these three areas- the reliability of a method, the reliability of its application, and the qualifications of individual witnesses).

237. For a more in-depth discussion and analysis of these claims and their subsequent treatment by courts *see generally* Gutierrez, *supra* note 24.

238. *Daubert*, 509 U.S. at 594.

239. *See e.g.*, PCAST Report, *supra* note 3, at 146; *but see* Cole & Sheck, *supra* note 91, at 810 ("an erroneous individualization is only one of many ways in which [forensic] error can be implicated in wrongful convictions"); Maria Cuellar, *The problem with eliminations: Why forensic comparisons need false negative rates*, 11 Forensic Sci. int'l: Synergy 100621, 1 (2025) (calling an exclusive focus on false positive rates "myopic" and calling for greater attention to false eliminations and other accuracy issues).

240. *See e.g.*, Monson *et al.*, *supra* note 194, at 1–3.

241. *See* Garrett et al., *supra* note 36, at 141–46; Gutierrez, *supra* note 24, at 1553–56.

242. *See e.g.*, Valerio & Bunn, *supra* note 127 ("When an examiner opines that a fired casing came from a particular firearm, they are accurate more than 99 percent of the time"); Agar, *supra* note 127, at 193 (providing same figure); Federal Bureau of Investigation,

statisticians outside the field have regarded these claims with nigh-unmitigated skepticism.[243] Indeed, one group of scholars has gone so far as to explain that even if we ignore the substantial flaws of studies that have produced industry-approved estimates of misidentification, appropriately calculated confidence intervals still likely stretch up into the double digits.[244]

With few exceptions, however, courts have remained largely unpersuaded by these critiques.[245] Many have cited to proficiency test results in support of categorizing error rates associated with firearms examination as low enough to favor admissibility,[246] and the echo-chamber-like treatment of forensic sciences by judges[247] has meant that even when courts do not mention such testing explicitly in support of admitting firearms examination evidence, they are often nevertheless relying on CTS results implicitly or unwittingly.[248] Indeed, even

supra note 24, at 3 ("studies produced low false positive rates of approximately 1% or less"); Grzybowski *et al.*, *supra* note 6, at 216 (using CTS data to support a false positive rate of .6% for firearms examination); Ronald G. Nichols, *supra* note 17, at 592 (placing the false positive rate for firearms examination at 1% based on CTS results); U.S. Dep't of Justice, *supra* note 99, at 23 (asserting that "[t]he aggregate false positive error rate in firearms/toolmarks studies to date falls well below" 5%).

243. Dorfman & Valiant, *supra* note 170, at 7 ("a decision of inconclusive in a non-test-blind study can mask what would be a mistaken identification. . . in casework"); *see also supra* Part II.A.

244. *See* Cuellar *et al.*, *supra* note 30, at 4–5.

245. *See e.g.*, 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. Tibbs, No. 2016-CF1–19431, 2019 WL 4359486, at *13–18 (D.C. Super. Ct. 2019) (citing the use of set-based studies as well as test-taking bias, and the issue of inconclusive conclusions as undermining the trustworthiness of false positive estimates for firearms examination); Abruquah, 296 A.3d at 991–992, (Md. 2023) ("As a second example, although advocates of firearms identification tout periodic proficiency testing by Collaborative Testing Services Inc. . . . as a method of ensuring the quality of firearms identification, the record contains no evidence supporting efficacy of that testing. To the contrary, the evidence suggests that examiners rarely, if ever, fail CTS proficiency tests. Dr. Hamby confirmed that the industry's mandate to CTS with respect to proficiency tests 'was to try to make them [as] inexpensive as possible.'")

246. *See e.g., United States v. Montiero*, 407 F.Supp.2d 351, 367–68 (D. Mass 2006); *United States v. Taylor*, 663 F. Supp. 2d 1170, 1177 (D.N.M. 2009); *United States v. Diaz*, No. CR 05–00167 WHA, 2007 U.S. Dist. LEXIS 13152, *26–28 (N.D. Cal. 2007); *United States v. Otero*, 849 F.Supp.2d 425, 433–434 (D.N.J. 2012); *United States v. McCluskey*, CR. No. 10–2734 JCH, 2013 U.S. Dist. LEXIS 203723, *24–27 (D.N.M. 2013); *United States v. Johnson*, (S5) 16 Cr. 281 (PGG), 2019 U.S. Dist. LEXIS 39590, *55–57 (S.D.N.Y. 2019); *United States v. Johnson*, No. 14-cr-00412-THE, 2015 U.S. Dist. LEXIS 111921, *10 (N.D. Cal. 2015); *United States v. Ashburn*, 88 F.Supp.3d 239, 246 (E.D.N.Y. 2015).

247. M. Chris Fabricant & Tucker Carrington, The Shifted Paradigm: Forensic Scince's Overdue Evolution from Magic to Law, 4 Va. J. Crim. L. 1, 37 (2010).

248. The court in *Otero*, for example, relied extensively and explicitly on CTS results to support admitting firearms examination evidence and concluding that the discipline is subject to only a low error rate hovering around 1%. 849 F.Supp.2d at 433–34. Some of the myriad courts that have echoed its reasoning have referenced proficiency testing explicitly as well. *See e.g., United States v. Lee*, 2022 U.S. Dist. LEXIS 150054, *7–8

as the scholarly community in dissent against firearms examination's validity has grown, judges have near-uniformly continued to accept industry-insider estimates of a 2%-or-less false positive rate.[249] And they have done so even after espousing critical views of existing tests,[250] as well as when classifying an approximately 2% false positive rate figure as weighing against admissibility.[251]

If judges desire to hold firearms examiners to Rule 702's requirement that expert testimony "help the trier of fact,"[252] then the acknowledgement by some that accuracy testing is "not meaningful when all of the *expert*'s peers always passed"[253] does not go far enough. It may be that the error rate studies relied on by courts don't just fail to sort good experts from bad, but cannot so much as distinguish experts from novices. Courts and commentators have explained Rule 702's helpfulness inquiry as centering on whether an expert's specialized

---

(N.D. Ill. Aug. 22, 2022); *Ashburn*, 88 F.Supp.3d at 246. But many others have relied on the *Otero* court's conclusion without directly referencing its basis in proficiency testing. *See e.g., United States v. Dunham*, 654 F.Supp.3d 1183, 1190–1191 (E.D. Okla. 2023); *State v. Raynor*, 254 A.3d 874, n.7 (Conn. 2020); *People v. Robinson*, 2013 IL App (1st) 102476, ¶¶ 85–87. The same goes for *Monteiro*, which relied on CTS results while also issuing caution about their simplicity, 407 F.Supp.2d at 367: multiple courts have relied on its ultimate decision to admit without acknowledging the role played by proficiency testing. *See e.g., State v. Fuentes*, 228 P.3d 1181, 1187 (N.M. Ct. App. 2009); *Turner v. State*, 953 N.E.2d 1039, 1052 (Ind. 2011); *Robinson*, 2013 IL App (1st) 102476, ¶¶ 85–87.

249. Gutierrez, *supra* note 24, at 1553–56 (collecting cases); *see also United States v. Richardson*, 2024 U.S. Dist. LEXIS 39167, *22–24 (D. Kan. Mar. 6, 2024). Indeed, some courts have been satisfied with false positive rates as high as 5%. *McCluskey*, 2013 U.S. Dist. LEXIS 203723 at *24–27.

250. The judge in *Montiero*, for example, caveated that error rates derived from CTS testing "may not stand up to scrutiny" and characterized them as potentially "somewhat elementary," but she nevertheless concluded that the "known error rate is not unacceptably high," and emphasized that the defense presented "no evidence that the tests are inaccurate or otherwise deficient." *Montiero*, 407 F.Supp.2d at 367–68. Professors Garrett and Mitchell have characterized similar statements made by a judge in the context of fingerprint examination as burden shifting. Garrett & Mitchell, *supra* note 43, at 933. But other judges have nevertheless followed suit. *See e.g., Taylor*, 663 F. Supp. 2d at 1177 (acknowledging "uneven test administration, make-up, and level of difficulty significantly limit the usefulness of . . . " CTS results, but concluding "[n]onetheless, this number [misidentification rates from CTS tests of around 1%] at least suggests that the error rate is quite low"); *United States v. Hunt*, 464 F. Supp. 3d 1252, 1258 (W.D. Okla. 2020) (reviewing studies on firearms examination's accuracy and concluding "the known error rate is not unacceptably high. Defendant does not introduce any contradictory studies") (internal quotations & citations omitted); *Johnson*, 2019 U.S. Dist. LEXIS 39590 at *56–57 (acknowledging doubts about the usefulness of existing error rate data on firearms examination but holding that "even accepting . . . that the error rate could be as high as 1 in 46, or close to 2.2%, such an error rate is not impermissibly high.").

251. *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, 1265 (D. Or. 2020).

252. Fed. R. Evid. 702(a); *compare* Ill. R. Evid. 702 (mandating that "scientific, technical, or other specialized knowledge will *assist* the trier of fact") (emphasis added).

253. *Montiero*, 407 F.Supp.2d at 367–68 (*quoting United States v. Lewis*, 220 F. Supp. 2d 548, 554 (S.D. W. Va. 2002) (emphasis added).

knowledge and skills fall "beyond the ken of the average layperson,"[254] such that they can address subjects outside the experience of jurors or shed additional light—through their "more thorough or refined understanding"—even on more common or familiar matters.[255] Firearms examination should struggle to navigate even this initial hurdle for admissibility given that, unlike other forensic fields,[256] it has failed to demonstrate that the training practitioners undergo actually allows them to compare fired bullets and cartridge cases more accurately than lay people.[257] It follows from the structure of a helpfulness requirement that studies into the error rate associated with firearms examination "cannot offer even minimally useful evidence of validity . . . unless capable of discriminating between expert and lay performance."[258] And at least one published example frequently cited by proponents of firearms examination and courts admitting testimony from the field[259] has already tripped at this low bar.[260] All told, then, courts have struck truer than they know by questioning whether the

254. *Girardot v. United States*, 92 A.3d 1107, 1109 (D.C. Ct. App. 2014).

255. J. Harvey Brown, *Eight Gates for Expert Witnesses*, 36 Hous. L. Rev. 743, 751–53 (1999); *see also* Fed. R. Evid. 702 advisory committee's notes ("Whether the situation is a proper one for the use of expert testimony is to be determined on the basis of assisting the trier. There is no more certain test for determining when experts may be used than the common sense inquiry whether the untrained layman would be qualified to determine intelligently and to the best possible degree the particular issue without enlightenment from those having a specialized understanding of the subject involved in the dispute.") (internal citations & quotations omitted); People v. King, 161 N.E.3d 143, 153 (Ill. 2020) (applying a similar test in a *Frye* state).

256. *See e.g.*, Tangen et al., *supra* note 156, at 997 (comparing expert and novice performance on three sets of prints and concluding: "We have shown that qualified, court-practicing fingerprint experts are exceedingly accurate compared with novices . . . Expertise with fingerprints appears to provide a real performance benefit, but fingerprint experts—like doctors and pilots—make mistakes that can put lives and livelihoods at risk); David White et al., *Perceptual Expertise in Forensic Facial Image Comparison*, 282 Proc. R. Soc. B 20151292, 5 (2015) (performing similar experiment for facial recognition and deterring that "rank order of identification performance in all six experimental tests placed examiners over controls and students"); Moshe Kam et al., *Writer Identification by Professional Document Examiners*, 42 J. Forensic. Sci. 778, 778 (1997) ("The results of our test lay to rest the debate over whether or not professional document examiners possess writer-identification skills absent in the general population. They do").

257. Gutierrez & Prokesch, *supra* note 13, at 1336 ("[N]o validation effort that we are aware of has so much as established the existence of expertise in the comparison of firearms evidence."). This would not preclude firearms examiners from offering more elementary or background information about guns. *See* Garrett & Mitchell, *supra* note 43, at 913 (discussing the role of experts merely providing information about general principles); *People v. Williams*, 2025 Colo. App. LEXIS 342 *15 (Colo. App. Ct. 2025) ("Testimony that a certain gun is *capable* of firing certain ammunition does not require similarly specialized knowledge or experience") (emphasis original).

258. Gutierrez & Prokesch, *supra* note 13, at 1336.

259. Duez et al., *supra* note 108, *cited by* Federal Bureau of Investigation, *supra* note 24, at 3 at Table 1; Department of Justice, *supra* note 99, at 23; *United States v. Harris*, 502 F. Supp. 3d 28, 37 (D.D.C. 2020).

260. Gutierrez & Prokesch, *supra* note 13, at 1334 (finding that the study used different source samples so simplistic that a post hoc control group of attorneys were able to perform at a level "essentially indistinguishable from that of trained examiners")

difficulty in proficiency and validation testing of firearms examination might limit what can be ascertained about the discipline's false positive rate. What they have lacked—at least by their own accounts—is access to data validating those fears and demonstrating the consequences of testing samples more reflective of the real world.

To be clear, even before CTS 23–5262 and even if putting aside criticisms of study design, judicial faith in a 1–2% false positive rate estimate could not withstand scrutiny. Most critically, and though defendants may not have raised these studies during challenges to the admissibility of firearms examination,[261] four validation efforts released prior to CTS 23–5262 had produced false positive rates well in excess of industry-preferred estimates (and exceeding even generous floors set for scientific validity)[262]: as high as 13.1%, 14.3%, 21.1%, or 39.6% when considering just comparisons in which examiners reached definitive source conclusions.[263] And those studies likely provide more reasonable estimates of the rate of misidentifications in casework because (1) examiners are likely to confront challenging comparisons in casework; (2) the studies eliciting double-digit misidentification rates account for at least some of those challenges; and (3) studies reflecting false positive rates in the low single digits, in contrast, failed to do so because they did not consciously seek out and include the most difficult available comparisons in their sample sets.[264]

But to the extent that previous examples have failed to reach courts (and at times, to sway them),[265] CTS 23–5262 should shatter whatever flimsy foundations remain for judicial confidence in single-digit false positive estimates for firearms examination. Not only did the misidentification rate it generated (as high as 42.2%) exceed typical court estimates by twenty times or more, but it also likely does a better job of replicating typical casework challenges. Critically, bullets will rarely reach firearms examiners in pristine condition; instead, contact with materials like "wood, metal, drywall, glass, . . . [or] bullet proof vest[s]," may damage fired bullets,[266] "distorting or even destroying the markings left by the barrel of the firearm and making comparisons difficult."[267] But rather than embrace samples typical of casework (and subject

261. *See, e.g., Montiero*, 407 F.Supp.2d at 368; *Hunt*, 464 F. Supp. 3d at 1258; *but see Winfield*, No. 15CR14066–01, at 24; Defendant's Post-Frye/Rule 403 Hearing Brief, *supra* note 69, at 17–18.

262. *See* PCAST Report, *supra* note 3, at 151–52.

263. *Supra* note 32.

264. *See* Gutierrez, *supra* note 24, at 1561–73.

265. *See e.g.*, United States v. Pete, No. 3:22cr48, 2023 U.S. Dist. LEXIS 135065, n.5 (N.D. Fl. July 31, 2023) (refusing to adopt defendant's argument for a higher false positive rate by questioning the "external validity" of a European study "to provide useful information about error rates for firearms and toolmark examination in the United States").

266. Best & Gardner, *supra* note 32, at 31 & 34 (2022) (further describing such damage as "common in casework" and emphasizing the need to include damaged specimens in validation testing so that comparisons "are representative of casework").

267. Brandon A. Best, *Thesis: An Assessment of the Foundational Validity of Firearms Identification Using Ten Consecutively Button-Rifled Barrels*, at iv (2020) (on file with author); Schwartz, *supra* note 17, at 11 (citing damage to firearms and fired ammunition as a primary source of difficulty for examiners).

to the kinds of weathering and abuse common in the field), most validation efforts have instead gone out of their way to ensure pristine comparison items: firing into specialized collection tanks or catches and even discarding items "deformed" in any way.[268] Although the producers of CTS 23–5262 may not have replicated casework conditions by firing bullets into hard materials, their inadvertent lapses in quality control nevertheless introduced bullets whose individual characteristics (already minimal) were subject to other types of damage (specifically, tank rash).[269] By including such realistic (and eminently fair) challenges, CTS 23–5262 did what few other studies have managed, and explored a far larger and more meaningful portion of "the full range and distribution of types and difficulty normally seen in casework."[270] Given that we know, from what little data exists, that error and damage are correlated,[271] the false positive rate generated therefore likely paints a more accurate picture of the chance of misidentification in casework.

Moreover, neither courts that have previously found firearms examination's false positive rate to favor admissibility, nor the field's proponents and practitioners, can easily disregard the results of CTS 23–5262. Although the latter have tended to simply omit at least the Knapp & Garvin, Pauw-Vugts *et al.*, and Mattijssen *et al.* studies from their summaries of foundational research,[272] what brief mentions they have made of them (only in response to defense arguments raising these studies rather than as a result of their own independent efforts to provide transparent and thorough accounts of error rate studies) have quickly dismissed the significance of higher false positive rates based on, among other issues, studies' use of casts and photos, inclusion of European examiners, and the differing conclusion scales they addressed.[273]

268. Monson et al., *supra* note 170, at 9; *see also* Guyll et al., *supra* note 98, at 2 ("Mesh cuboid-shaped cartridge-case catchers caught the fired cartridge-cases as they were ejected from the firearm. All cartridges cases that came to rest anywhere other than the inside of the catcher were discarded").

269. *Supra* notes 218–20.

270. *Supra* notes 92 & 106.

271. *See, e.g.*, Best, *supra* note 267, at iv ("[D]amaged bullets have almost twice as much error and uncertainty of measurement when compared to undamaged bullets.").

272. *See, e.g.*, FBI, *supra* note 24, at 18–20; U.S. Dep't of Justice, *supra* note 99, at 23–24; Organization of Scientific Area Committees 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 [https://perma.cc/XVH6–7HWM]; Trans. of Proceedings, Illinois v. Winfield, No. 15CR14066–01, at 242–43 (Cir. Ct. Cook Cnty. Mar. 1, 2022) (Todd Weller, a firearms examiner, admitted to eliding mention of existing studies with higher false positive rates multiple times during prior sworn testimony).

273. FBI, *supra* note 24, at n.13 ("noting that a group of statisticians had cited to three European error rate studies that yielded a higher false positive rate" but dismissing those studies because "[t]wo . . . used a verbal scale which did not employ definitive statements" and the third involved "looking at cast replicas"); Trans. of Proceedings, *Illinois v. Winfield*, 15CR14066–01, at 245–47 (criticizing the generalizability of Mattijssen et al. because participants "were looking at static photographs of comparisons and explaining prior failures to mention the study during testimony as follows: "So one reason why I'm not presented that study is because I don't think it is a test of the methodology where you are looking at physical samples under the

Such distinctions ring suspicious and hypocritical at best,[274] but they have, at times, persuaded courts.[275] They apply, however, with even less force to CTS 23–5262, which used bullet samples rather than reproductions, had examiners respond in line with the AFTE Range of Conclusions (identification, inconclusive, or elimination), and produced a substantial false positive rate (as high as 25.6%) even among participants from accredited laboratories.[276] The remaining reasons to believe that CTS 23–5262 may not reasonably estimate the potential for error in casework[277] not only undercut studies favored by proponents of firearms examination[278]—they also tend to suggest that CTS 23–5262 may indeed *underestimate* the chance of a misidentification in the real world.[279]

Perhaps more simply still, having for so long (and so frequently) relied on CTS proficiency testing results despite their deficiencies,[280] judicial, practitioner, and law enforcement defenders of firearms examination have no legitimate, internally consistent grounds on which to ignore the near-chance performance of participants from CTS 23–5262. This has not stopped defenders of the field from speaking out of both sides of their mouths.[281] But as the old adage on geese and ganders suggests, if those tests sufficed to support faith in *de minimis* risk of misidentification, then so too must they serve to shatter assumptions about the discipline's validity when practitioners so widely fail to perform up to snuff. Though courts should never have demanded that

---

microscope, and it is using a different means of conclusions").

274. Gutierrez, *supra* 24, at 1572 ("[L]aw 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.").
275. *See e.g., Pete*, 2023 U.S. Dist. LEXIS 135065, at n.5; *but see Winfield*, No. 15CR14066–01 at 13 (accepting higher false positive rate estimates and likening the experience of learning about them to touring a "basement room of horrors").
276. *Supra* note 214 & Part II.
277. Among other issues, participants knew they were being tested and 24% of them failed to return responses to Collaborative Testing Services (a rate of attrition, by the corporation's own admission, "higher than other firearms texts over the past five years"). *Supra* note 192.
278. *See e.g.*, Monson et al., *supra* note 170, at 5 & 11 (noting that test was declared and that only 79 of 270 participants completed the full study, an attrition rate of 70.7%); Best & Gardner, *supra* note 32, at 34 (acknowledging the test was declared an that only 44 of 66 participants completed the study, an attrition rate of 33.3%); Baldwin et al., *supra* note 78, at 3 (test was declared and was completed by only 218 out of 284 participants, an attrition rate of 23.2%).
279. *See e.g.*, Scurich et al., *supra* note 92, at 1336; William C. Thompson, *Shifting decision thresholds can undermine the probative value and legal utility of forensic pattern-matching evidence*, 120 120 Proc. Nat'l Academy Scis. e2301844120 (2023).
280. *Supra* note 274.
281. *Contrast* Carroll et al., *supra* note 28, at 142 ("the PTRC [which included firearms examiner Todd Weller] strongly recommends against calculating an overall false positive error rate from this test as being reflective of the discipline as a whole or reflective of any individual examiner"); *with* Weller, *supra* note 201, at 24–25 ("Another measure of professional error can be gleamed from examining proficiency test records" and providing single digit false positive rates.).

challengers demonstrate an invalidating misidentification rate for the field of firearms examination,[282] the accessibility of CTS 23–5262[283] and the emphasis placed on appropriate burdens of proof by recent amendments to Rule 702[284] combined should lead to more skepticism and less reflexive treatment of bullet comparison evidence such that future reviews cannot so easily string cite judicial uniformity behind purely aspirational false positive rates.

B. *Validity As Applied for Damaged Bullets*

Even if courts refuse to reassess their adherence to the dogma of 1–2% false positive rates for firearms examination to reject the field's "foundational validity," i.e., its admissibility under Rule 702(c),[285] the results of CTS 23–5262 should nevertheless reverberate in admissibility decisions specifically dealing with conditions mirroring that test: comparisons performed on damaged bullets. Rule 702(d) directs courts to evaluate whether an "expert's opinion reflects a reliable application of the principles and methods to the facts of the case," which scientific scholars have termed "validity as applied."[286] And Rule 403's[287] balancing of unfair prejudice against probative value provides similar recourse, both in addition to Rule 702(d) as well as in jurisdictions that do not employ a similar admissibility test.[288] Indeed, not only have scientific schol-

282. *Supra* note 250.

283. The results of CTS 23–5262 are available publicly and without charge on Collaborative Testing Service's website. *Supra* note 138. But judges and litigants would struggle to gain access to some of the other articles relied on by proponents of the field or which produced double-digit false positive rates. While three have appeared in open access journals or under open access licenses, *see* Monson et al., *supra* note 194; Guyll *et al.*, *supra* note 98; Baldwin et al., *supra* note 78, others can only be reviewed with a journal subscription, *see* Mattijssen *et al.*, *supra* note 32, or have appeared in either the AFTE Journal or as presentations at an AFTE conference, *see* Pauw-Vugts *et al.*, *supra* note 32; Knapp & Garvin, *supra* note 32; Best & Gardner, *supra* note 32. At least one court has heavily criticized the AFTE Journal for its lack of accessibility. *United States v. Tibbs,* No. 2016 CF1 19431, 2019 D.C. Super. LEXIS 9*,* at *30–32 (D.C. Super. Ct. 2019) ("AFTE does not make this publication generally available to the public or to the world of possible reviewers and commentators outside of the organization's membership. Of course, an interested party can receive the publication by joining AFTE, if such a person meets the organization's membership requirements, or can pay to access specific articles. But unlike other scientific journals, the AFTE Journal is not more broadly available and cannot even be obtained in university libraries . . . [t]he AFTE Journal, by generally limiting the review of its publications and making them available only to its members or others who pay, avoids the scrutiny of scientists and academics outside the field of firearms and toolmark analysis"). And interested parties can locate the conference presentation only by searching AFTE's internal index of DVDs and paying the organization $40. *See* Ass'n Firearms & Toolmark Exam'rs, *Training Seminar Presentation Index*, https://afte.org/resources/training-seminar-presentation-index/ [https://perma.cc/8JEC-UYY6] (last visited May 22, 2026).

284. *Supra* note 232.

285. PCAST Report, *supra* note 3, at 4–5.

286. *Id.* at 5; *see also* Imwinkelreid, *supra* note 235, at 3–4.

287. Fed. R. Evid. 403.

288. *See, e.g., Daubert*, 509 U.S. at 595 ("[e]xpert evidence can be both powerful and quite misleading because of the difficulty in evaluating it. Because of this risk, the judge . . . under Rule 403 of the present rules exercises more control over experts than over lay

ars followed this approach, taking even valid methods to task when experts practice outside established parameters for reliable application,[289] but courts have followed suit and excluded narrower applications or subtypes of evidence from multiple fields otherwise and more generally deemed admissible.[290] And though several such narrower challenges have failed in the specific context of firearms and toolmark examination,[291] at least a few notable decisions may light the path forward.

To begin, in *Ramirez v. State*, the Florida Supreme Court ruled inadmissible under a *Frye* general acceptance standard the testimony of a forensic scientist purporting to identify a knife found in the defendant's possession as the "particular knife to the exclusion of all others" responsible for a murder victim's stab wound.[292] Although the court recognized "[t]he theory underlying tool mark evidence" broadly as "generally accepted in the scientific community and . . . long . . . upheld by courts," it also emphasized that the testimony at issue in the particular case it reviewed (involving comparison of marks left in human cartilage rather than more deeply studied materials like wood and metal) "depart[ed] from traditional knife mark identification theory in significant ways."[293] After concluding that the level of peer review, existence of blind testing, and knowledge about the error rate for identifying the source weapon responsible for stabbing marks in human cartilage compared unfavorably with

witnesses) (internal quotations & citations omitted); People v. McKown, 924 N.E.2d 941, 956 (Ill. 2010) ("This balancing of risk of prejudice versus probative value is not performed across the board. It is necessarily a case-by-case analysis. Our finding that HGN evidence meets the *Frye* standard does not preclude the possibility that, in a given case, the trial court might rule such evidence inadmissible on grounds of undue prejudice.").

289. PCAST Report, *supra* note 3, at 82 (opining that DNA analysis of simple mixtures is foundationally valid, but that existing studies limited the validity of analysis of complex mixtures to "mixtures of 3 individuals in which the minor contributor constitutes at least 20 percent of the intact DNA" and cautioning that while "[t]he range in which foundational validity has been established is likely to grow as adequate evidence for more complex mixtures is obtained and published . . . validity as applied . . . [requires] a special emphasis on ensuring that the method was applied correctly and within its empirically established range").

290. People v. Ferguson, 526 N.E.2d 525, 530–532 (Ill. App. Ct. 2d Dist. 1988) (acknowledging the general acceptance of shoeprint identification but excluding the testimony of an expert who claimed to be able to identify a suspect based on wear patterns repeated across multiple pairs of his shoes); Murray v. Motorola, Inc., 2014 D.C. Super. LEXIS 16, 49 (D.C. Super. Ct. 2014) ("the evidence and testimony at the *Frye* hearing show that what [the expert] actually did to reach his conclusions was so far removed from what is generally accepted as true WOE analysis that his methodology amounts to little more than a subjective section of sources"); State v. McPhaul, 808 S.E.2d 294, 305 (N.C. Ct. App. 2017) (excluding the testimony of a fingerprint examiner and because he "implicitly asked the jury to accept her expert opinion that the prints matched" without any supporting testimony or documentation of the similarities observed); United States v. Williams, 382 F. Supp. 3d 928, 938, (N.D. Cal. 2019) (excluding results of a probabilistic genotyping system when applied to a mixture more complex than the system had been validated to handle).

291. *See* Garrett et al., *supra* note 36.

292. Ramirez v. State, 810 So.2d 836, 841 (Fl. 2001).

293. *Id.* at 845–47.

the same factors supporting the admissibility of toolmark evidence more generally,[294] the court emphasized that the proponent of scientific evidence needed to prove the general acceptance of "the testing procedures used to apply the principle to the facts of the case at hand" and reversed the defendant's conviction as based on "unreliable and inadmissible" expert testimony.[295]

Two courts have reached markedly similar conclusions regarding comparison of magazine marks.[296] In *Sexton v. State*, the Texas Court of Criminal Appeals evaluated a firearms examiner's claim that live rounds of ammunition discovered at the defendant's home had definitively (i.e., a conclusion reached "with one hundred percent certainty") been cycled through the same magazine as spent cartridge cases linked to a shooting.[297] Finding literature supportive, at best, only of an examiner's ability to opine on the source of magazine marks when in possession of "knowledge of the manufacturing process of the tool surface and . . . the tool available for creating test toolmarks" (both of which the examiner in *Sexton* lacked), the court ruled that while "the underlying theory of toolmark examination could be reliable in a given case . . . the State failed to produce evidence of the reliability of the technique used in this case . . . "[298] And in *State v. Kuek*, a Nebraska trial court went even further, excluding an examiner's opinion about magazine marks—under both the *Daubert* standard and a balancing of probative value versus prejudice—despite the examiner possessing the suspect firearm and magazine.[299] Even though Nebraska courts had long

294. *Id.* at 849–51.

295. *Id.* at 851. Indeed, though *Ramirez* singled out knife marks in cartilage, far less research and no pairwise / sample-to-sample error rate studies exist for non-firearm toolmarks (think screwdrivers, knives, pliers, etc . . .), leading some scholars to opine that "[t]here is every reason to think that those other types of toolmark comparison raise similar or far larger reliability concerns" compared with cartridge case and bullet comparison evidence. Garrett et al., *supra* note 36, at 158; *see also* People v. Genrich, 572 P.3d 604, 610, 613 (Colo. App. 2025) (affirming grant of new trial during postconviction proceedings based on new evidence regarding the scientific community's rejection of toolmark comparison evidence).

296. Sexton v. State, 93 S.W.3d 96, 101 (Tex. Ct. Crim. App. 2002); State v. Kuek, No. CR 19–2918 (Neb. Dist. Ct. Dec. 4, 2023). In the case of semiautomatic handguns, cartridges are held in a container "which has a spring and a follower to feed those cartridges into the chamber." Glossary, *supra* note 73, at 72. The lip of that container "may scrape the sides of each case as they are pushed into a chamber, or as they are loaded into, or removed from, the magazine by hand," even if "the magazine is unattached to the firearm," leaving behind magazine marks. Thompson, *supra* note 69, at 22; *see also* Glossary, *supra* note 73, at 73; NRC Report, *supra* note 4, at 45. But though firearms examiners consider such marks an appropriate basis for identification conclusions, Thompson, *supra* note 69, at 2, they rely on and study them with less frequency that markings on areas like a breech face or firing pin both over concern regarding subclass characteristics, and (perhaps more importantly) because they are technically unrelated to the firing process (i.e. they can support a conclusion only that cartridges were loaded into or cycled through the same gun, not that those cartridges were fired from the same gun). Ronald Nichols, *Subclass Characteristics: From Origin to Evaluation*, 50 AFTE J. 68, 82 & 84 (2018); Glossary, *supra* note 73, at 43.

297. Sexton, 93 S.W.3d at 99.

298. *Id.* at 101.

299. *Kuek*, No. CR 19–2918, at 4, 10, 13.

admitted evidence from the firearm and toolmark examination field generally,[300] the trial judge held "that precedential backdrop . . . [was] outweighed" because "[w]hile [the examiner in its case] engaged in feature comparison, it was not the time-tested type of feature comparison typically seen in criminal cases."[301] Instead, the court found that, whereas "a stack of research studies" supported toolmark examination generally, "[n]one of those studies deal[t] with an examiner directly making the mark," (as was true in the case under review).[302] In other words, "the battle to be had over forensic toolmark examinations" could be left "for another day" because the examiner's more particularized methodology relying on magazine marks was simply "too untested for [the] Court to find it valid."[303]

Indeed, an Illinois trial court in *People v. Kimberley* even recently relied centrally on proficiency testing to support a case-specific exclusion of cartridge case comparison testimony.[304] Faced with a broad challenge to the admissibility of firearms examination evidence outright, as well as arguments directed at the specific application of firearms examination methods by the Illinois State Police, the court "rejected Kimberley's request for a Rule 702/*Frye* hearing because the general admissibility of firearm identification testimony [was] settled law in Illinois."[305] The court nevertheless heard significant testimony critical of the accuracy and consistency of firearms examination writ large and reached multiple conclusions about the field as a whole, for example that error rates had not been appropriately "quantified" because of the failure to blind participants as well as the reality that "error rates in many studies . . . are wildly varying . . . [a]t their best . . . almost 0%, while at their worst . . . closer to 66%."[306] But it also clarified that the hearing it allowed and the ruling it reached were "not about problems with firearms identification generally."[307] Instead, the court emphasized that the "hearing addressed problems *specific to this case*," going on to discuss, among other things, the ways firearms examiners at the Illinois State Police encounter task-irrelevant,[308] biasing information that might skew their conclusions and "*most importantly that the examiner*

300. *Id.* at 10, *citing State v. Wheeler*, 956 N.W.2d 708 (Neb. 2021); *State v. Mason*, 709 N.W.2d 638 (Neb. 2006); *State v. Lopez*, 544 N.W.2d 845 (Neb. 1996); *State v. Hill*, 851 N.W.2d 670 (Neb. 2014).

301. *Kuek*, No. CR 19–2918 at 10.

302. *Id.* at 9.

303. *Id.* at 13.

304. *Kimberley*, 22-CR-02932–01, at 57–59.

305. *Id.* at 57, *citing* People v. Rodriguez, 141379-B (Ill. App. 1st. 2018).

306. *Kimberley*, 22-CR-02932–01, at 43–44, 46.

307. *Id.* at 57.

308. Nat'l Comm'n Forensic Sci., Views of the Commission Ensuring That Forensic Analysis Is Based Upon Task-Relevant Information 3–5 (2015) (warning that bias from task irrelevant information could corrupt forensic conclusions and providing the following definition: "Information is *task-irrelevant* if it is not necessary for drawing conclusions about the propositions in question, if it assists only in drawing conclusions from something other than the physical evidence designated for examination, or if it assists only in drawing conclusions by some means other than an appropriate analytic method"), https://www.justice.gov/ncfs/file/818196/dl?inline= [https://perma.cc/3UKP-D43H].

. . . failed to make an elimination determination during an external proficiency test . . . [with] no evidence suggest[ing] any corrective measures were taken."[309] In other words, though the judge in *Kimberley* felt constrained in his ability to translate broadside criticisms of firearms examination and the "wishful thinking" of its proponents about "minimal" error rates into action, Rule 403's flexible balancing test and the litigation of case-specific flaws in the analysis conducted at the Illinois State Police (especially a proficiency testing failure) freed him to more narrowly point out that the accuracy of the analyst involved "could be worse than that of her colleagues," to worry over whether even excellent trial advocacy could effectively confront firearms examination evidence, and ultimately to conclude that "no less restrictive remedies" other than a case-specific exclusion could address the undue prejudice that would result from allowing an "expert" to testify regarding the source of crime scene cartridge cases.[310]

Litigants facing firearms examination should take note of the decisions in *Ramirez*, *Sexton*, *Kuek*, and *Kimberley*; seek out ways the analysis performed in their case might depart from best practices or from the existing research into accuracy; and weave those narrower criticisms into their admissibility challenges. But beyond merely *informing* the practice of defense attorneys, those decisions should also produce results and should give other judges the confidence to more surgically single out testimony concerning the source of damaged bullets for exclusion. Specifically, *Kimberley*'s analysis would support a court in excluding the testimony of an individual analyst who failed CTS 23–5262 and yet would seek to be qualified as capable of accurately comparing other-than-pristine bullets (more on this shortly).[311] But it is *Ramirez*, *Sexton*, and *Kuek* that provide a model for more impactful and sweeping exclusions. Historically, firearms and toolmark examiners cited monolithically to accuracy studies (in other words, cited to the total available record of accuracy studies without regard to the extent that any specific type of toolmark or sample had featured).[312] But more recently, even discipline insiders have begun to recognize that because accuracy "varie[s] notably among firearm : ammunition combinations," no one should "assum[e] that examiner performance results for a given firearm : ammunition combination are necessarily applicable to a different firearm : ammunition combination."[313] And it is this elementary and yet somehow-belated acknowledgment—when combined with exhortations in rules of evidence to evaluate the reliability of a method in application as well as assess the threat of undue prejudice posed—that should ultimately herald the

309. *Kimberley*, 22-CR-02932–01, at 57.

310. *Id.* at 56–57.

311. *Infra* Part III.C.

312. *See, e.g.*, Sci. Working Group for Firearms & Toolmarks, Systemic Requirements/Recommendations for the Forensic Firearm and Toolmark Laboratory 1 (2010) ("The reliability of the science has been demonstrated and supported through proficiency tests and validity studies over many decades. The calculated error rates indicate that the conclusions reached are accurate when appropriate methods are followed by a competent examiner.").

313. Bullet Black Box Working Grp., *supra* note 117, at 4.

doom of source conclusion testimony based on comparison of damaged bullets. To state it bluntly, like the knife and magazine mark evidence just discussed, precious little evidence exists about the accuracy of damaged-bullet comparisons, leaving CTS 23–5262's disturbing and near-chance misidentification rate as the only real data point.

Problematic and insufficient though it might be,[314] the library of accuracy research concerning firearms examination skews heavily toward the testing of cartridge case comparisons.[315] Of the studies performed to date that allow false positive rate calculations,[316] eight have exclusively involved cartridge cases,[317] compared to just four involving bullets.[318] Especially because damage does not as heavily impact the comparison of cartridge cases, the former provide a poor basis from which to estimate the accuracy and potential for error when examiners confront deformed bullets.[319] But the latter provide an almost equally poor supplement. The only study involving a sizable number of participants (173) used exclusively pristine bullets.[320] Another, although including damaged samples, specifically cautioned that its results, because of its small number of participants (forty-nine), should be considered only "exploratory, hypothesis generating."[321] And the last two (only one of which included damaged bullets), not only similarly involved glaringly small numbers of participants (forty-four and sixty-four) but also produced far-from-comforting or validating estimates of accuracy and error (false positive rates as high as 21.1% and 14.5% and specificity as low as 21% and 58%).[322] If, as firearms examiners concede, damage

314. *Supra* Part I.A.
315. *See* Hicklin *et al.*, *supra* note 113, at 2 ("Most of the firearm studies conducted to date have been limited to cartridge cases, not bullets").
316. Bullet Black Box Working Grp., *supra* note 117.
317. *See* Baldwin et al., *supra* note 78, at 3; Guyll et al., *supra* note 98, at 1; Keisler et al., *supra* note 97, at 56; Duez et al., *supra* note 108, at 1073; Chapnick et al., *supra* note 110, at 557; Law & Morris, *supra* note 190, at 1704; Mattijssen *et al.*, *supra* note 32, at 2; Erwin J. A. T. Mattijssen et al., *Firearm examination: Examiner judgments and computer-based Comparisons*, 66 J. Forensic Sci. 96, 96 (2021).
318. *See* Hicklin et al., *supra* note 113, at 1; Best & Gardner, *supra* note 32, at 32; Pauw-Vugts et al., *supra* note 32, at 115; Monson et al., *supra* note 194, at 86.
319. *See* Thompson, *supra* note 69, at 26 (noting damage is especially relevant to the availability of marks on bullets); Trans. of Proceedings, *Illinois v. Winfield*, 15CR14066–01, at 59–61 (Cir. Ct. Cook Cnty. 2022) (Research scientist and former laboratory director Jeff Salyards testified that including damaged bullets in validation studies is critical and also distinguished cartridge case research: "so cartridge cases, right. It's a little more straightforward than bullets. They—it's rare that in a forensic setting we would get half a cartridge case, right, or that it would be really severely deformed") (on file with author).
320. Monson et al., *supra* note 170, at 9, 11 ("Bullets that became deformed from impacting the walls of the water tank were not used in this study"); *see also* Hicklin et al., *supra* note 113, at 13 (noting that Monson *et al.* included only undamaged bullets).
321. Hicklin et al., *supra* note 113, at 12–13.
322. The false positive rates noted are the 95% upper bound of Clopper Pearson exact confidence intervals. *See* Best & Gardner, *supra* note 32, at 33–35 (providing participant and performance figures and explaining that one third of the bullets in each test set were deliberately damaged); Pauw-Vugts et al., *supra* note 32, at 115, 125 (providing participant and performance figures and making no mention of damage to

significantly impacts accuracy,[323] then the evidence (outside of CTS 23–5262) relevant to the validity of firearms examination methods when applied to the types of bullets most commonly encountered in casework[324] is therefore at best almost entirely absent, and at worst, deeply troubling and outside scientific expectations of validity.[325] It ultimately leaves CTS 23–5262 as the *only* test of a significant number of examiners (around 280)[326] to involve damaged bullets, and thus to provide anything even remotely resembling an estimate of accuracy and error we might expect on such samples in casework.[327] Unless judges would give examiners a pass even for scarcely edging out the performance of flipping a coin, they should therefore now exclude opinions about the source of deformed bullets under Rule 702(d), Rule 403, or their state equivalents.

C. *Separating the Wheat from the Chaff & Exploring Qualifications At & Before Trial*

It seems likely—given judicial reticence in criminal cases to exclude evidence offered by prosecutors[328]—that most litigants ultimately will see their challenges to firearms examination methods fall on deaf ears. Nevertheless, they must persevere and push for CTS 23–5262 to play a sizable role in the

bullet samples).

323. *Supra* note 267; *see also* Bullet Black Box Working Grp., *supra* note 117, at 5, ("*lower-quality bullets (i.e., bullets with more damage) tend to have higher rates of inconclusive responses*" & "When citing decision rates . . . the citation should refer to the quality (extent of damage) of the bullets being compared. Decision rates are affected by bullet quality.").
324. *Supra* note 221.
325. *See* PCAST Report, *supra* note 3, at 151–52.
326. *Supra* Part III.
327. It is surely true that no single study could ever validate a methodology and demonstrate conclusively its accuracy and consistency. *See e.g.*, Cuellar et al., *supra* note 30, at 13 ("at least two studies are generally necessary . . . But we emphasize that the studies must have adequate quality in their design and analysis . . . Furthermore, scientific validity depends not only on adequate study design and analysis, but also on the results"). Firearm examination insiders could be correct that the results of CTS 23–5262 cannot define their field as a whole. *See* Carroll et al., *supra* note 28, at 142. And admittedly, CTS 23–5262 limitations would give even objective outside researchers pause in over-interpreting its performance measures. *Supra* note 114. But none of that should save the admissibility of damaged bullet comparison testimony. I would argue differently had the test produced favorable estimates of accuracy and error; in other words, if the *proponents* of expert testimony about damaged bullets (who would bear the burden of proof under Rule 702) were to rely on only a single test to demonstrate validity. *Supra* note 258. But those challenging admissibility have no such burden. And it is hard to conceive of better evidence *rebutting* claims to the validity of bullet comparison evidence than CTS 23–5262.
328. *See also* Erin Murphy, *Neuroscience and the Civil/Criminal Daubert Divide*, 85 Fordham L. Rev. 619, 621–24 (2016) (documenting disparate outcomes under *Daubert*, success for civil defendants and prosecutors and widespread rejections for civil plaintiffs and criminal defendants); Peter J. Neufeld, *The (Near) Irrelevance of Daubert to Criminal Justice and Some Suggestions for Reform*, 95 Am. J. Pub. Health S107, S110 (2005) (same); *see generally* 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).

admissibility and treatment at trial of comparison testimony offered by *examiners who failed* that test. As in *Kimberley*, that might occur in the sense that an examiner's inability to navigate the challenges of CTS 23–5262 could provide the final straw for a successful as-applied challenge under Rules 702(d) or 403.[329] But even before reaching the reliability of scientific and technical evidence, Rule 702 and its state equivalents require a threshold showing that any witness who intends to provide opinion testimony about such methods be "qualified as an expert by knowledge, skill, experience, training, or education," such that their "specialized knowledge will help the trier of fact."[330] As Professors Garrett and Mitchell have argued eloquently and at length, the social science research presented above showing the ways that surrogates for empirically derived measures of performance are at best imperfect indicators of expertise[331] strongly supports the adoption of a "proficiency-based approach to expert qualification." Indeed, even though Rule 702 likely ought to be revised to bring its language unequivocally in line with scientific understandings on what qualifies someone as an expert,[332] its current text nevertheless supports a "performance-based interpretation" predicated on requiring "*demonstrable* expertise" based on empirically derived and individual evidence of accuracy (i.e., proficiency testing).[333] As they put it, "an . . . interpretation that is consistent with the text and advisory note, and that has more merit from a functional perspective, is that Rule 702 cares more about expertise per se than about any particular route to expertise," and thus, courts should focus on whether "the proffered witness is capable of expert performance outside the courtroom on the kinds of questions the expert will be asked to answer in the courtroom."[334]

Unfortunately, prior to CTS 23–5262, litigants facing off against firearms examiners were unlikely to derive much benefit from Garrett and Mitchell's

329. *Supra* note 113.
330. Fed. R. Evid. 702; Ill. R. Evid. 702.
331. *Supra* notes 48–61, 85.
332. Garrett & Mitchell, *supra* note 43, at 910; *see also* Imwinkelried, *supra* note 235, at 6.
333. Garrett & Mitchell, *supra* note 43, at 910; *see also* Imwinkelried, *supra* note 235, at 6 (describing Rule 702's existing language as "inapt" for Garrett & Mitchell's approach because it "mixes apples and oranges: the essence of the expertise (knowledge, skill, or a combination of knowledge and skill) versus the methods of acquiring the expertise (experience, training, or education)"). I, like Garrett and Mitchell, would support further amendments to Rule 702 that clarify that experience, training, and education, those "methods of acquiring" just mentioned, *id.*, must actually result in knowledge and skills beyond that of the average layperson and thus helpful to the trier of fact. *Accord* Watkins v. Schmitt, 172 Ill. 2d 193, 205 (1996) (expert testimony must involve "knowledge and application of principles of science beyond the ken of the average juror") (internal quotation & citation omitted); Brown, *supra* note 255 ("If the fact-finder is equally competent to examine an issue, the court will strike the expert's opinion under [a] helpfulness standard"). I also note, however, that some courts have also reached the common-sense conclusion that even a flexible qualification inquiry does not mean that any and all forms of training or experience would suffice. *See e.g.*, People v. Hunley, 313 Ill. App. 3d 16, 30 (1st Dist. 2000) (the "degree of knowledge or experience necessary to qualify the witness as an expert depends on the complexity of the subject matter and the likelihood of error if the witness is not properly qualified").
334. Garrett & Mitchell, *supra* note 43, at 912–13.

arguments. Because even an empirical view of Rule 702 might require only a showing that a witness could "perform better than chance and better than a layperson would,"[335] the number of examiners participating in routine proficiency testing and the ease of those tests combined would have rendered adoption of their approach little more than empirical "window dressing."[336] Litigants were simply unlikely to encounter analysts who had failed any proficiency test (at least in the sense of committing a false positive),[337] and though the available literature might alert them to the existence of other examiners who failed to navigate more formal validation studies, they would likely be unable to ever identify whether they had encountered such a practitioner.[338] Indeed, though appellate cases grappling with challenges to the qualifications of expert witnesses are closer to exception than norm,[339] and some of that minimal judicial record does evince a *near total* disregard for the relevance of proficiency testing information to admissibility,[340] concerns about whether an increased focus on existing proficiency tests would actually improve the quality of forensic evidence allowed into our criminal legal system are quite well-founded. It is actually common to find decisions in which firearms examiners have cited to their own proficiency testing results to buttress their qualifications, and judges have routinely and reflexively accepted such testimony when granting examiners the status of expert witnesses.[341] For essentially the first time, however, CTS 23–5262 and the substantial number of examiners who failed to surmount its challenges have offered up a chance for litigants to intervene and attempt to push Professors Garrett and Mitchell's arguments from proposal to reality. Doing so will require that litigants (1) manage to obtain access to individual

335. *Id.* at 923; *see also* Imwinkelried, *supra* note 235, at 12–13 ("[T]he most sensible standard is whether the record demonstrates that the witness has sufficient skill to reach the correct outcome more frequently than random chance and more often than the lay jurors.").
336. Max et al., *supra* note 136, at 294.
337. *Supra* Table II.
338. *Infra* notes 343–47.
339. Garrett & Mitchell, *supra* note 43, at 935; *see also* Imwinkelried, *supra* note 235, at 7.
340. *See, e.g.*, Ricks v. Pauch, No. 17–12784, 2020 LX 50109, at *40–48 (E.D. Mich. Mar. 23, 2020) (qualifying two long-retired firearms examiners as experts despite complete lack of proficiency testing); United States v. Hunt, 464 F. Supp. 3d 1252, 1260 (W.D. Okla. 2020) (holding that Rule 702 does not require rigorous proficiency testing).
341. *See, e.g.*, United States v. Boone, No. 23 Cr. 427 (PGG), 2024 U.S. Dist. LEXIS 89710, at *7–8, *17–18 (S.D.N.Y. May 13, 2024); United States v. Davis, No.: 4:18-cr-00011, 2019 U.S. Dist. LEXIS 155037, at *17–18 (W.D. Va. Sep. 11, 2019); United States v. Gist-Holden, 629 F. Supp. 3d 841, 845–46 (N.D. Ind. 2022); *Willie v. State*, 274 So. 3d 934, 940 (Miss. Ct. App. 2018); State v. Wheeler, 956 N.W.2d 708, 721 (Neb. 2021). Indeed, courts have even seen (and sometimes accepted) attempts by firearms examiners to use their personal performance on proficiency testing to claim an individual error rate lower than ones seen in discipline-wide studies. *See* State v. Miller, 852 S.E.2d 704, 715 (N.C. Ct. App. 2020) (Zachary, J., dissenting) (firearms examiner used proficiency test results to claim a lower false positive rate); Morones v. State, No. 09–16–00317-CR, 2017 Tex. App. LEXIS 10964, at *8 (Tex. Crim. App. Nov. 22, 2017) (firearms examiner claimed "[t]here is no real known rate of error . . . As far as my errors, I have never made an error on a proficiency test . . . my rate would be zero.").

examiner results from that test and (2) convince judges to take examiner failures seriously, either by refusing to qualify deficient examiners as experts or by allowing attorneys leeway to inquire into proficiency testing performance during trials. The law on both fronts may be murky to say the least, but it is also true that compelling evidence specific to the available record on the accuracy and persuasiveness of firearms examination testimony falls strongly behind access to individualized results from CTS 23–5262 and the potential of those results to make criminal trials fairer.

As an initial matter, even gaining access to whether a particular examiner passed or failed past exercises assessing their accuracy (proficiency testing or validation studies) may prove daunting for litigants. Though CTS makes Summary Reports publicly available online, those records do not contain identifiable participant information; instead, litigants must use the discovery process to link the individualized web codes from a Summary Report back to the examiner involved in their case.[342] And that lack of transparency unfortunately mirrors approaches taken more widely across research into the accuracy of firearms examination in which study authors have relied on protections for human research subjects[343] to justify shielding participant information not just from the public, but from defendants and even participating examiners themselves. One validation study[344] "both warned participants against checking their own results and used a process that would avoid a digital paper trail even for those participants who wished to track their performance."[345] And the author of a second research effort[346] candidly responded to an examiner's request for their own performance data during a webinar by explaining:

> It's a great idea and I applaud the person who asked the question for doing so because that certainly indicates a desire to improve. From our perspective, we did not allow individual participants to become aware of their results because we did not want that to become *Giglio* material in the event that they did.[347]

In other words, the researcher knew the Constitution mandates the disclosure of exculpatory or impeaching information,[348] recognized poor performance in his study would likely qualify and therefore disregarded the wishes of one of

342. *Supra* notes 137–39.

343. *See generally* 28 C.F.R. § 46 (2026); 45 C.F.R. § 46.116 (2026) (codifying protections for human research subjects).

344. Hicklin *et al.*, *supra* note 113.

345. Emily J. Prokesch & Richard E. Gutierrez, *Commentary on "R. Austin Hicklin et al. Accuracy and Reproducibility of Bullet Comparison Decisions by Forensic Examiners"*, 376 Forensic Sci. Int'l 112590, 3 (2025).

346. Monson et al., *supra* note 194.

347. Nat'l Inst. of Jus., *FBI Laboratory Decision Analysis Studies in Pattern Evidence Examinations*, at 49:31–50:01 (YouTube, Sep. 19, 2025), https://youtu.be/trvZ81u2g4?si=qDjJF_SESESIxY7u [https://perma.cc/AF3S-ANR4].

348. Brady v. Maryland, 373 U.S. 83, 87 (1963) ("[S]uppression by the prosecution of evidence favorable to an accused upon request violates due process where the evidence is material either to guilt or to punishment, irrespective of the good faith or bad faith of the prosecution."); Giglio v. United States, 405 U.S. 150, 154 (1972) (same for impeachment information, i.e. "evidence affecting credibility").

his volunteers to assess and improve their own performance all to undermine the rights of criminal defendants.

Some of that secrecy may be eroding, at least in federal laboratories.[349] But if prosecutors or labs will not voluntarily provide information about an examiner's past performance on tests, litigants representing criminal defendants will need to try to compel this discovery. Problematically, however, whether they should have access to proficiency testing data at all is far from established. The Federal Rules of Criminal Procedure do not mention the disclosure of proficiency testing results explicitly,[350] nor do the discovery rules for criminal cases in forty-eight of the fifty states.[351] Only Massachusetts and New York[352] have gone a different route, mandating that the prosecution turn over to the defense "[d]escriptions of negative outcomes of proficiency testing . . . of the expert witness or of any testing or laboratory facility used by the expert for tests or experimentation"[353] and "a list of proficiency tests and results administered or taken within the past ten years."[354] Things are equally mixed in the realm of caselaw. Decisions reviewing requests for discovery on proficiency testing are rare, "[b]ut one still finds cases in which the courts conclude that proficiency

349. *See* Letter from Elizabeth J. Shapiro, Deputy Dir., U.S. Dep't. of Just., to U.S. Dist. Ct. J. Patrick Schiltz, Possible Amendment to Federal Rule of Evidence 702, at 6 (Nov. 6, 2020), https://www.uscourts.gov/forms-rules/records-rules-committees/suggestions/department-justice-20-ev-aa [https://perma.cc/P52L-799G] ("The Department recognizes that a forensic examiner's past performance on relevant, skill-based testing is an important measure for evaluating her performance in a given case. As such, FBI proficiency test results are routinely provided to defense counsel upon request. The FBI Laboratory will soon begin disclosing proficiency test results without a specific defense request as part of their general discovery and disclosure procedures").

350. Fed. R. Crim. P. 16(a)(1)(G)(iii) (listing required disclosures for expert witnesses as "a complete statement of all opinions . . . the bases and reasons for them; the witness's qualifications, including a list of all publications authored in the previous 10 years; and a list of all other cases in which, during the previous 4 years, the witness has testified . . . ").

351. *See, e.g.*, Cal. Penal Code § 1054.1(f) (Deering 2026) (requiring disclosure only of "any reports or statements of experts made in conjunction with the case, including the results of physical or mental examinations, scientific tests, experiments, or comparisons which the prosecutor intends to offer in evidence at the trial"); Me. R. Crim. P. 16(a)(2)(G) (2026) ("Any reports or statements of experts, made in connection with the particular case, including results of physical or mental examinations and of scientific tests, experiments, or comparisons"); Ky. R. Crim. P. 7.24(1)(c) (2026) ("upon written request by the defense, the attorney for the Commonwealth shall furnish to the defendant a written summary of any expert testimony that the Commonwealth intends to introduce at trial. This summary must identify the witness and describe the witness's opinions, the bases and reasons for those opinions, and the witness's qualifications").

352. Illinois comes close: for expert testimony generally, it requires disclosure only of "any reports or statements of experts, made in connection with the particular case, including results of physical or mental examinations and of scientific tests, experiments, or comparisons, and a statement of qualifications of the expert," *see* Ill. Sup. Ct. R. 412(a)(iv) (2026), but it does mention disclosure specifically of "proficiency testing results" when the prosecution is the proponent of DNA evidence, *see* Ill. Sup. Ct. R. 417(b)(v) (2026).

353. Mass. R. Crim. P. 14(b)(2)(C)(iii)(b).

354. N.Y. Crim. Proc. Law § 245.20(1)(f) (McKinney 2025).

data is not material to a criminal defendant's defense or of sufficient relevance to compel discovery in civil cases."[355] On the other hand, multiple courts across the country have recently characterized proficiency testing errors as material impeachment and thus discoverable (even if harmless error analysis has sometimes made these rulings more moral victory than anything else for specific criminal defendants).[356]

Table IV: Poor Performers in Firearms Examination Blackbox Studies[357]

Study	Bullets or Cartridge Cases	% / # Examiners	FPR PCAST
Baldwin et al.	CC	2% / 5	40%
Law & Morris	CC	11% / 2	40%
Monson et al.	CC	10% / 18	17.4%
Monson et al.	B	6% / 10	21.3%
Best & Gardner	B	2% / 1	50%
Guyll et al.	CC	2% / 5	27.8%
Hicklin et al.	B	18% / 9	18.2%

It therefore appears likely that litigants out to procure individualized proficiency testing data will need to meticulously demonstrate its materiality and potential exculpatory value to courts. And luckily here, the state of research into the accuracy of firearms examination strongly supports such requests as well as Garrett and Mitchell's general contention that "[j]urors should have information about the actual performance of an expert and not just their professional pedigree."[358] As I have argued above, pedigrees (training, experience, degrees, etc . . .) are poor surrogates for expertise in the form of superior (and thus helpful to jurors) performance.[359] But in the firearms examination realm, not only is it true that no study has ever confirmed a relationship between training or experience and superior performance, but also that "available data

355. Garrett & Mitchell, *supra* note 43, at 945.
356. *See, e.g.*, *id.* at 946 (collecting cases); United States v. Shipp, 422 F. Supp. 3d 762, 778 n.11 (describing proficiency testing results as "pertinent for the determining whether an expert "reliably applied the principles and methods to the facts of the case"); Commonwealth v. Sullivan, 85 N.E.3d 934, 945 (Mass. 2017) ("Although evidence of the criminalist's failure to pass proficiency tests was, contrary to the motion judge's ruling, both exculpatory and admissible, its nondisclosure was not sufficiently prejudicial to warrant a new trial").
357. This table pulls from all the pairwise studies that have made individualized data available. *See* Baldwin et al., *supra* note 78; Law & Morris, *supra* note 190; Monson et al., *supra* note 194; Best & Gardner, *supra* note 32; Guyll *et al.*, *supra* note 98; Hicklin *et al.*, *supra* note 113. I have already explained the calculations behind most in prior publications. *See* Gutierrez, *supra* note 31; Gutierrez & Addyman, *supra* note 31; Prokesch & Gutierrez, *supra* note 345. Data from the last, *see* Best & Gardner, *supra* note 43, is available from the thesis of one of its authors and reveals that the single examiner who reached a false positive only eliminated on one other within class comparison. *See* Best, *supra* note 267, at 101–02.
358. Garrett & Mitchell, *supra* note 43, at 946.
359. *Supra* Section I.

troubles rather than confirms an association . . . not only can novices perform just as well as professional examiners on at least certain comparisons; but prior black box studies have consistently failed to isolate any participant characteristics predictive of examiner proclivity for committing false positive errors."[360] Indeed, as shown in Table IV above, poor performing examiners (i.e., examiners who commit false positives) have appeared across multiple studies of firearms examination's accuracy.

That alone "should spark concern, not because every member of every profession can be expected to perform at the highest level at all times, but because such a sizable minority of practitioners committing false positives at double-digit or near double-digit rates simply cannot be reconciled with discipline claims regarding the practical impossibility of error."[361] But, even more critically, that data pertains to questions of the materiality, exculpatory value, and impeachment value of individualized performance evidence, because none of those studies have managed in any way to link superior or deficient performance to the surrogates for expertise so beloved by Rule 702 and courts, like training, experience, and laboratory accreditation.[362] In other words, in the context of firearms examination, litigants have access to hard and fast proof that as judges and juries evaluate the qualifications and credibility of proposed experts, resorts to training and experience may only lead them astray.[363] Instead, the only thing that might reveal an examiner who "excels at . . . serving primarily the needs of the prosecution . . . at the expense of misidentifications"[364] and thereby impeach sweeping claims about expertise made during voir dire is data from rigorous tests of performance. As the court in *State v. Adams* recognized, triers of fact cannot be left "lack[ing] information that would allow them to determine where on the error spectrum the testifying examiner falls."[365] Only access to the individualized results of rigorous tests like CTS 23–5262 can remedy that conundrum.

The same data on poor performers also lends strong support to Garrett and Mitchell's more significant call for a performance, as opposed to pedigree, approach to qualifying witnesses as experts in the arena of firearms examination.[366] Unless courts violate canons of interpretation and treat Rule 702's language about "help[ing] the trier of fact" as mere "surplusage,"[367] that

360. Prokesch & Gutierrez, *supra* note 345, at 1.

361. *Id.* at 2.

362. *Id.* at 1; Gutierrez, *supra* note 31, at 3 (triers of fact are left in the "unenviable position of being required to vet the qualifications of firearms examination witnesses, knowing performance varies widely and into double-digit rates of misidentification . . . and yet bereft of tools to sort experts from charlatans").

363. Indeed, in many cases the worst performing examiners are among the most experienced. *See* Prokesch & Gutierrez, *supra* note 345, at 1; Law & Morris, *supra* note 190, at 1714.

364. Prokesch Gutierrez, *supra* note 345, at 3.

365. State v. Adams, 340 Or. App. 661, 699 (Or. Ct. App. 2025).

366. *Supra* notes 331–34.

367. Obduskey v. McCarthy & Holthus LLP, 586 U.S. 466, 476 (2019) ("explaining that courts "generally presume that statutes do not contain surplusage" and seek to "giv[e] effect to every word") (internal quotations & citations omitted).

field's failure to establish a connection between the Rule's various categories ("knowledge, skill, experience, training, or education") and accuracy means that prosecutors cannot meet their burden to establish expertise without reference to a witness's record of passing rigorous tests. In other words, prosecutors cannot show that a witness is sufficiently helpful unless that witness is accurate. And a witness cannot show they are accurate if they have either failed past tests or have never participated in a challenging one. Indeed, at least one court has taken this approach, refusing to qualify a fingerprint examiner as an expert witness despite her decade of experience, eighteen months of training, and prior qualification as an expert by other courts because she had more recently failed proficiency tests.[368] But if courts follow the science and follow applicable principles of interpretation, then litigants facing those examiners who either failed CTS 23–5262 or never took that test should be able to successfully bar those witnesses from testifying as experts. That said, because Garrett and Mitchell have already made the case for an empirically driven Rule 702 at length, because even deficient examiners can dodge these arguments by engaging in corrective actions or retraining with their laboratory,[369] and because of the prevailing judicial view that Rule 702's qualification requirements set an almost nonexistent bar,[370] this Article will not linger further in the realm of judicial gatekeeping. Rather, it will land, at last, where litigants likely will: smack dab in the middle of trials involving firearms examiners.

In such a final stand in defense of their clients, litigants can, however paradoxically, take some solace in having lost the battle over qualifications. After all, judges who have refused to buy in to arguments about proficiency testing and admissibility have often done so by resort to the refrain that such issues "go to . . . weight . . . but not . . . admissibility."[371] To deprive attorneys of the opportunity to raise proficiency testing errors during cross-examination following such statements would therefore require judges to speak from both sides of their mouths (and not very subtly). More importantly though, whatever a judge decides regarding a witness's ability to pass Rule 702's threshold qualification criteria, the jury remains the ultimate arbiter of credibility. As a Ninth Circuit model jury instruction, for example, notes: "opinion testimony should be judged like any other testimony. [Jurors] may accept it or reject it

368. United States v. Cloud, 576 F. Supp. 3d 827, 847 (E.D. Wash. 2021) ("the Court cannot in good conscience qualify (fingerprint examiner) Watts as an expert with the requisite skill to perform fingerprint comparisons when her two most recent proficiency exams either contained an error or required a significant amount of assistance from her supervisor"). *But see* United States v. Laraba, No. 19-CR-36-RJA, 2023 U.S. Dist. LEXIS 115039, at *6–10 (W.D.N.Y. July 5, 2023) (distinguishing *Cloud* and qualifying DNA analyst where that analyst had passed all proficiency tests).

369. *See* Eldridge et al., *supra* note 142, at 3–5 (describing common laboratory approaches after an error as including retraining and testing).

370. *Supra* note 335.

371. J.H. v. Cruz, No. 3:14-cv-02356, 2023 U.S. Dist. LEXIS 14244, at *22–23 (M.D. Tenn. Jan. 27, 2023); *see also* Harell v. State, No. 1109, 2023 Md. App. LEXIS 757, at *31 (Md. App. Ct. Nov. 7, 2023) (defendant's "challenges to . . . Special Agent Swick's proficiency . . . went to the weight, rather than the admissibility").

and give it as much weight as [they] think it deserves."[372] Indeed, the U.S. Supreme Court has even refused to "relax the requirements of the Confrontation Clause" when it comes to expert witnesses (including those "possess[ing] the scientific acumen of Mme. Curie and the veracity of Mother Theresa"), instead emphasizing the critical ways cross-examination might "weed out not only the fraudulent analyst, but the incompetent one as well."[373]

Proficiency testing errors can "counter" proponents' efforts to "create the impression that a witness has developed skill in applying a methodology by repeatedly—and successfully—conducting analysis employing it" as well as more generally "she[d] light on the degree of the witness's expertise."[374] But more than that, studies into lay perceptions of firearms examination testimony show they are absolutely essential to those aims (and thus to the fairness of trials involving bullet and cartridge case comparison testimony). In line with the longstanding judicial warnings mentioned earlier about the "talismanic significance"[375] jurors *accord* scientific evidence and their "natural inclination . . . to equate science with truth,"[376] it appears that laypeople "assume a high degree of validity in firearms comparison and are likely to convict based on such evidence if left unrebutted."[377] In fact, most estimate the field's "false positive rate to be orders of magnitude smaller" than the very best results from validation studies.[378] Worse, because of the discipline's systemic bias against eliminations, most defendants (even innocent ones) will lack the means to counter these unwarranted impressions and push jurors to decide guilt and innocence accurately and fairly within the limits of the science: they will not find a firearms examiner to reach a contrary (and correct) elimination, the independent scientists they might hire will not replicate the exculpatory impact of such a missing rebuttal, and common testimonial limits favored by judges are unlikely to have any effect.[379] Cross-examination has its own limitations. But studies show that "evidence-based" approaches can get the job done and course correct the undue willingness of jurors to defer blindly to the conclusions of

372. Manual of Model Crim. Jury Instructions for the Dist. Cts. of the 9th Circ. 55 (9th Cir. Jury Instructions Comm. 2025)), https://www.ce9.uscourts.gov/jury-instructions/model-criminal [https://perma.cc/4DGN-A5PG]; *see also* United States v. Rutland, 372 F.3d 543, 546 (3d Cir. 2004) ("It is not improper for jurors to consider an expert's experience and credentials when determining the weight of the expert's testimony"); Dixon v. Int'l Harvester Co., 754 F.2d 573, 580 (5th Cir. 1985) (once a witness has been "properly admitted as an expert, the jury [is] at liberty to accept or reject his testimony, and to judge his credibility").

373. Melendez-Diaz v. Massachusetts, 557 U.S. 305, 319, 319 n.6, 325 (2009).

374. Imwinkelried, *supra* note 235, at 21, 31.

375. United States v. Frazier, 387 F.3d 1244, 1263 (11th Cir. 2004).

376. *In re* Det. of New, 21 N.E.3d 406, 412 (Ill. 2014)

377. Garrett et al., *The Impact of Defense Experts on Juror Perceptions of Firearms Examination Testimony, supra* note 15, at 241.

378. *Id.* at 234, 241.

379. *See id.* at 241–45; Nicholas Scurich et al., *Ineffectiveness of the "Consistent With" Judicial Limitation on Forensic Firearm Identification Testimony*, 49 L. & Human Behav. 387, 395 (2025); Garrett et al., *Mock Jurors' Evaluation of Firearm Examiner Testimony, supra* note 36, at 422.

firearms examiners.[380] One has even focused on proficiency testing specifically to conclude that it "informed participants' opinions of the testifying examiner's convincingness, skill, competence, and confidence."[381] For all the myriad reasons I have laid out, CTS 23–5262 can and should push firearms examination (especially of damaged bullets) out of the courts before it ever reaches jurors. But if it does not, judges must at least refrain from tying defendants' hands behind their backs. If, in other words, CTS 23–5262 cannot ripple in admissibility decisions, then (if defendants are to have their guilt weighed fairly) at least its deficient participants should feel its sting during cross-examination.

CONCLUSION

While we will never know how many defendants like Patrick Pursley firearms examination has helped falsely accuse and incarcerate, we can be certain both that he has not suffered alone[382] and that the scope of requests for firearms analysis ensures that in thousands of cases each year[383] tragedy may strike

380. *See* William C. Thompson & Nicholas Scurich, *How Cross-Examination on Subjectivity and Bias Affects Jurors' Evaluations of Forensic Science Evidence*, 64 J. FORENSIC SCI. 1379, 1379 (2019).

381. William E. Crozier et al., *Juror Appraisals of Forensic Evidence: Effects of Blind Proficiency and Cross-Examination*, 315 FORENSIC SCI. INT'L 110433, 9 (2020).

382. *See* Gutierrez, *supra* note 24, at 1537 (cataloging the wrongful arrest and prosecution of at least five other men due to firearms examination errors); MICHIGAN STATE POLICE FORENSIC SCI. DIV., AUDIT OF THE DETROIT POLICE DEPARTMENT FORENSIC SERVICES LABORATORY FIREARMS UNIT 5 (2008), https://www.sado.org/articles/Article/58 [https://perma.cc/HWA8-EWWV] (24 misidentifications discovered during the audit of the Detroit Police Department were uncovered after reviewing just 283 total cases, for a real-world false positive rate of no less than 8.5%). Indeed, known false arrests and convictions likely but scratch the surface. *See, e.g.*, Cole, *supra* note 136, at 126–27; *see also* Andre A. Moenssens, *Novel Scientific Evidence in Criminal Cases: Some Words of Caution*, 84 J. CRIM. L. & CRIMINOLOGY 1, 12–13 (1993) (noting that misidentifications occur but "mistakes of this kind are not very likely to be discovered"). And firearms examiners admit as much, acknowledging the existence of a plethora of largely undocumented and untraceable misidentifications. *See* Bruce Moran, *A Report on the AFTE Theory of Identification & Range of Conclusions for Tool Mark Identification & Resulting Approaches to Casework*, 34 AFTE J. 227, 227 (2002) ("In the 1980s some striated toolmark misidentifications resulting from a poor understanding of toolmark criteria for identification were experienced. An increasing need to address problems of applying subjective criteria became apparent."); Evan E. Hodge, *Guarding Against Error*, 20 AFTE J. 290, 291–92 (1988) (noting that "most of us [firearms examiners] know someone who has committed serious error" and describing misidentification by another examiner of the wrong .45 caliber firearm due to cognitive bias and pressure from prosecutors); Alfred Biasotti & John Murdock, *The Scientific Basis of Firearms & Toolmark Identification*, MOD. SCI. EVIDENCE: L. & SCI. EXPERT TESTIMONY 143 (1997) (acknowledging misidentifications stemming "from one examiner ascribing too much significance to a small amount of matching striae and not appreciating that such agreement is achievable in known non-match comparisons"); Lowell Bradford, *Forensic Firearms Identification: Competence or Incompetence*, 11(2) AFTE J (1979) ("An appalling number of misidentifications have been found in the firearm identification field").

383. *See* MATTHEW R. DUROSE ET AL., BUREAU OF JUST. STATS., PUBLICLY FUNDED FORENSIC CRIME LABORATORIES: RESOURCES & SERVICES, 2014, at 3 tbl. 4 (2016), https://bjs.ojp.

again. Courts can and should put an end to that risk. The harrowing likelihood of misidentification brought to light by CTS 23–5262 justifies banishing the discipline from our criminal legal system. But for judges who may balk at ushering in such a sea change, it is worth considering that more rigorous proficiency testing could well have cut these injustices off at the pass by identifying examiners too willing to reach identifications based on limited or misleading markings. After all, damage and minimal markings provoked the errors seen both in Mr. Pursley's case and in CTS 23–5262.[384] If our criminal court system will not part with firearms "experts" but nevertheless hopes to prevent additional wrongful arrests, prosecutions, and convictions stemming from the faulty opinions of examiners, it will at minimum need far better access to data about examiner performance—about accuracy for the field at large as well as the limits of its application and those actually qualified to navigate source conclusions about fired ammunition.

CTS 23–5262, although still deeply imperfect, could light that narrower path forward. By using realistic and challenging samples, the test has already substantially enhanced our understanding of the practice of firearms examination. It provides the best available data on the difficulty posed by damage to bullets.[385] It lends support to the notion that accreditation acts as an effective quality control (though not a cure-all for what ails the discipline).[386] And it has helped cull the wheat from the chaff by identifying at least sixty-one poor performers[387] and by distinguishing between higher and lower quality laboratories. Texas can now tout its performance (though only so far given abysmal specificity),[388] while Wyoming finds itself awash in problems: its lab has been

gov/library/publications/publicly-funded-forensic-crime-laboratories-resources-and-services-2014 [https://perma.cc/92QH-W8AU]; Gun Violence in America, NAT'L INST. JUST. (Feb. 26, 2019), https://www.nij.gov/topics/crime/gun-violence/pages/welcome.aspx [https://perma.cc/Q2AW-6KBN].

384. *Supra* note 15; *see also* Garrett et al., *supra* note 36, at 153 (arguing that "more careful attention" to failed proficiency tests may have allowed D.C.'s crime lab to address deficiencies before they became severe enough to shutter the lab); *Merritt v. Arizona*, 425 F. Supp. 3d 1201, 1208–09 (D. Ariz. 2019) (discussing opinion of consulting firearms examiner that damage and minimal markings provoked the original firearms conclusions that led to the wrongful arrest of Leslie Merritt); Nicole Garcia, *Evidence Unsealed in Freeway Shooter Case*, Fox 10 Phoenix (May 12, 2016, 18:36 MT), https://www.fox10phoenix.com/news/evidence-unsealed-in-freeway-shooter-case [https://perma.cc/XBF9-LS66] (same); Megan Cassidy, *A Look Inside the Reports that Unraveled the Phoenix Freeway-Shooting Case*, Ariz. Republic (May 13, 2016, 7:26 MT), https://www.azcentral.com/story/news/local/phoenix/2016/05/12/ballistics-report-phoenix-freeway-shootings/84281444/ [https://perma.cc/W66T-SLWJ] (same). Two other independent firearms examiners would eventually agree with Mr. Haag. *Merritt v. Arizona*, No. CV-17–04540-PHX-DGC, 2020 U.S. Dist. LEXIS 152011, at *14–15 (D. Az. Aug. 21, 2020).

385. *Supra* Section III.B.

386. *Supra* note 214.

387. *Supra* Table III.

388. *See* Tex. Forensic Sci. Comm'n, Final Report on Complaint No. 21.27; National Innocence Project, University of Colorado Law School (Houston Police Dept Crime Lab; Firearms / Tool Marks) 36 (2024), https://www.txcourts.gov/fsc/news/complaint-2127-firearms-report-released/ [https://perma.cc/2LBZ-F4FP].

shuttered, and the state has been forced to spend money on outside contractors, and has been mired in review of over 400 past criminal cases.[389] In fact, at an even more granular level, post hoc review of the test and surveying its participants revealed the dangerous influence of cognitive bias and expectation, the insufficiency of nonblind verification, and even core misunderstandings of the "sufficient agreement" standard.[390] But without concerted action by courts, litigants, and critics, CTS 23–5262 will remain an outlier, proficiency test providers will revert to pleasing their purchasers with elementary comparisons, and courts will be left wanting for the data so essential to a robust review of firearms examination methods.

This is more than speculation; there is already precedent for such a retreat (and signs in the present that history will indeed repeat itself). Back in 1995, CTS administered[391] "an unusually difficult inter-laboratory comparison,"[392] involving prints from identical twins left in blood, mirroring some of the most significant challenges that might confront latent print examiners.[393] The results were, unsurprisingly, "nothing like the universal passage" discussed throughout this Article:[394] 22% of participants committed at least one misidentification.[395] Almost immediately—and beyond even just the value of confirming concerns about similarities between prints from twins and fears about the difficulty of comparing bloody prints—the test proved its worth, "spurr[ing] on discussion and action in the latent print community, including movements to develop national guidelines for training and quality assurance."[396] But just as quickly, the industry of latent print examiners reacted in "shock" and with "disbelief."[397] And CTS, it appears,

389. *See* Clair McFarland, *Wyoming Crime Lab Fails Proficiency, Has To Review More Than 400 Past Gun Tests*, COWBOY STATE DAILY, Sept. 20, 2024, https://cowboystatedaily.com/2024/09/20/wyoming-crime-lab-fails-proficiency-has-to-review-more-than-400-past-gun-tests/ [https://perma.cc/NA6K-BKRT]; Sarah Elmquist Squires, *State Firearms Lab Fails Accreditation*, THE RANGER, Sept. 20. 2024, https://rivertonranger.com/news/state/state-firearms-lab-fails-accreditation/article_13667ef4–779b-11ef-be59–7f52e933c638.html [https://perma.cc/36VM-6CXT]; Angus M. Thuermer Jr., *Wyoming Crime Lab problem prompts review of gun-involved prosecutions*, WYOFILE, Jul. 30, 2024, https://wyofile.com/wyoming-crime-lab-problem-prompts-review-of-gun-involved-prosecutions [https://perma.cc/NGE5-FXTT].

390. *See* Carroll et al., *supra* note 28, at 134, 141–42.

391. *See* COLLABORATIVE TESTING SERVS., INC., *Forensic Testing Program, Latent Prints Examination*, REPORT NO. 9508 (1995).

392. EXPERT WORKING GRP. ON HUM. FACTORS IN LATENT PRINT ANALYSIS, NAT'L INST. STANDARDS & TECH., LATENT PRINT EXAMINATION AND HUMAN FACTORS: IMPROVING THE PRACTICE THROUGH A SYSTEMS APPROACH 1, 6 (2012) (also providing images of the prints); *See also* Garrett & Mitchell, *supra* note 43, at 925 (describing the test as "too challenging for the participants); Koehler, *supra* note 66, at 1093 (calling the test "more difficult"); Cole, *supra* note 136, at 1030 (calling the test "notorious").

393. *See* Max et al., *supra* note 136, at 293; *See generally* Glenn Langenburg, *Deposition of Bloody Friction Ridge Impressions*, 58 J. FORENSIC IDENTIFICATION 355 (2008) (prints in blood); Xunqiang Tao, Xinjian Chen, Xin Yang, Jie Tian, *Fingerprint Recognition with Identical Twin Fingerprints*, 7 PLOS ONE e35704 1 (2012) (twins).

394. Max et al., *supra* note 136, at 293.

395. *See* Garrett & Mitchell, *supra* note 43, at 925.

396. Max et al., *supra* note 136, at 293.

397. David L. Grieve, *Possession of Truth*, 46 J. FORENSIC IDENTIFICATION 521, 524 (1996);

picked its purse over progress: never again demanding that latent print examiners pass through such a daunting gauntlet, and instead producing more simplistic tests that paved the way for vanishingly small false positive rates.[398]

In the wake of CTS 23–5262, much the same trajectory seems afoot. In multiple statements, CTS has affirmed its commitment to placating firearms examiners, saying that it hopes to "better meet the needs of the firearms community"[399] while reassuring examiners that it remains "committed to collaboration with and support of the communities that we serve."[400] And lest readers dismiss such statements as mere public relations veneer, empirical evidence also supports an interpretation of CTS's statements as harbingers of a less-than-rigorous future of testing; neither of the two exams released (as of this Article) since CTS 23–5262 involves any legitimate challenge whatsoever to examiners' abilities to perform accurately on elimination samples. The first, a bullet comparison proficiency test, included two different-source comparisons, but used guns from different manufacturers with distinguishable class characteristics (indeed examiners could have reached eliminations blindfolded and merely by touch).[401] Unsurprisingly, it provoked (at most) a single misidentification for a false positive rate of .1% and a specificity rate of 99.9%.[402] The second (a cartridge case proficiency test) was no less laughably easy, provoking no misidentifications and only a few inconclusives for a specificity rate of 98.2%.[403] Once again, the inclusion of guns from different manufacturers

Cf. Cole, *supra* note 136, at 1213 (noting the community "reacted with some consternation").

398. *See* Koehler, *supra* note 66, at 1093 ("[S]ubsequent tests have been less difficult in that they rarely include latents and knowns from such highly related individuals."); Garrett & Mitchell, *supra* note 43, at 926–29 (describing subsequent tests by saying they "usually provide quite clear latent prints, of questionable resemblance to the often-distorted latent prints collected from real crime scenes," cataloguing lower error rates across all latent print proficiency tests from CTS from 1996–2016, and concluding that their "results reinforce the concerns of other scholars who have . . . noted the . . . simplicity of the test designs"); Lyn Haber & Ralph Norman Haber, *Error Rates for Human Latent Fingerprint Examiners*, *in* Automatic Fingerprint Recognition Systems 339, 347 (Nalini K. Ratha & Ruud Bolle, eds., 2004) (analyzing CTS latent print proficiency tests from 1995–2001 and concluding that "number of respondents who achieve perfect performance increases, and the number of erroneous identifications remains small" after 1997); Cole, *supra* note 136, at 1213 ("results have improved since [1995]. On subsequent tests, the number of examiners reporting false positive results has ranged from 6% to less than 1 %. Similarly, the number of false positives as a proportion of the total comparisons has dropped below 1%"); Cole, *supra* note 136, at 1030 ("Since the notorious 1995 test, there has been a decline in false positives").

399. Collaborative Testing Servs., Inc., *supra* note 23 (frequently asked questions at 5).

400. Christopher Czyryca, *Letter: Collaborative Testing Services AFTE PTRC Report Statement* 2 (Collaborative Testing Servs. 2024).

401. *See* Collaborative Testing Servs., Inc., Firearms Examination Test No. 24–5261 2, 14 (2024) (the elimination samples came from a Hi-Point as opposed to Tangfolio used to fire the known bullets, they displayed eight lands and grooves as opposed to six).

402. *See id.* at 11 (the above caveats that the test provoked "at most" one false positive because that error appears to have been clerical; the true false positive rate is likely 0% and the true specificity rate likely 100%); *See id.* at 18, 68.

403. *See* Collaborative Testing Servs., Inc., Firearms Examination Test No. 24–5262 10

bearing distinguishable class characteristics stacked the deck in favor of accuracy.[404] Firearms examiners might revel in these results, but they do nothing for the courts, the accused, and the criminal legal system writ large.

That dismal, data-less future, however, is far from inevitable; critics, courts, and controversy could all play a role in forcing CTS's hand (and that of firearms examiners) on an arc towards more informative proficiency exercises.[405] The FBI claimed for years that its latent print examiners had never and would never err, that when they reached an identification it conclusively meant that "the suspect's finger is the only finger in the world, living or dead, human or primate, which could have produced the latent impression."[406] But the storm of criticism and controversy which engulfed the agency after misidentifying an American citizen, Brandon Mayfield, as the source of a fingerprint linked to a terrorist bombing in Madrid, forced it to change its tune and led to more tempered conclusions and more robust procedures for tackling cognitive bias.[407] One isolated exclusion decision based on skipping over lab-specific internal validation[408] pushed the developers of a system for interpreting complex DNA mixtures to conduct additional research[409] and to change their case consultation policies.[410] And the PCAST Report's takedown of set-based studies[411] ushered in a sea change: only one unpublished study allowing for the calculation of false positive rates existed before that report,[412] but more appropriate sample-to-sample designs have now become the norm.[413]

That we know anything at all about the risk of misidentification posed by firearms examination evidence thus stems, not from self-motivated evolution

---

(2025).

404. *See id.* at 2, 11, 13.

405. *See e.g.*, Koehler, *supra* note 66, at 1093 ("Because the forensic science community is unlikely to embrace and voluntarily participate in tough, realistic proficiency tests, the courts might *require* such participation from laboratories that wish to offer results in a trial. This point arose in the context of DNA evidence as well").

406. Cole, *supra* note 136, at 1196.

407. *See generally* Off. Insp. Gen, U.S. Dep't Just., A Review of the FBI's Progress in Responding to the Recommendations in the Office of the Inspector General Report on the Fingerprint Misidentification in the Brandon Mayfield Case 1, 16 (2011).

408. *See People v. Hillary*, 2015–15 (N.Y. Sup. Ct. Aug. 26, 2016) (otherwise, evidence from the system has been uniformly admitted); *See e.g., United States v. Lewis*, 442 F. Supp. 3d 1122 (D. Minn. 2020); *See also United States v. Christensen*, 2019 U.S. Dist. LEXIS 24623 *1, *6 (C.D. Ill. Feb. 15, 2019).

409. *See* Hannah Kelly et al., *A Sensitivity Analysis to Determine the Robustness of STRmix™ with Respect to Laboratory Calibration*, 35 Forensic Sci. Int'l: Genetics 113 (2018).

410. *See* Trans. Of Proceedings, *Illinois v. Corey Morgan, Dwight Doty, & Kevin Edwards*, Nos. 16 CR 08715–01, 16 CR 08715–02, & 16 CR 08715–03, at 83–86 & 101–02 (Cir. Dt. Cook Cnty. June 18, 2019) (system developer discussing change to assisting laboratories complete validation before consulting) (on file with author).

411. PCAST Report, *supra* note 3, at 106–11.

412. *Id.* at 111.

413. *See* Guyll et al., *supra* note 98, at 2, 9; Keisler et al., *supra* note 97, at 56; Monson et al., *supra* note 194, at 88; *See also* Jaimie A. Smith, *Beretta barrel fired bullet validation study*, 66 J. Forensic Sci. 547 (2020).

on the part of practitioners, but from outside pressure.[414] But courts cannot stop now, cannot allow the field to rest on its rotten laurels. Mark Twain once wrote, "the weakest of all weak things is a virtue which has not been tested in the fire."[415] As with virtue, so must it be for bullet and cartridge case comparisons. If firearms examiners are to deserve a privileged seat as experts, if courts are to fulfill their critical role as gatekeepers, and if scientific consensus is ever to develop around the validity of tracing bullets and cartridges back to a source weapon, CTS 23–5262 cannot stand alone. Far more data culled from realistic, challenging comparisons is necessary to either prove the field valid or fully substantiate the fears of its critics. If courts heed the analysis in this Article and begin excluding firearms examination evidence (*in toto* or with a more targeted focus on damaged bullets and deficient examiners), we may well get there. Faced with judges who consistently refuse to accept vanishingly low false-positive estimates based on softball studies, firearms examiners will have to load up with others utilizing the worst-of-the-worst, the most difficult of comparisons, and hope for better results. Seeing bullet comparison evidence excluded outright given the dearth of essentially any comforting data, firearms examiners could only move forward by conducting additional studies specifically targeting damaged bullets. And if firearms examiners come to expect that they will be denied a seat in the witness stand without proof of having passed meaningful proficiency tests, we might expect them to begin demanding, not the simplistic fodder they have historically favored, but tests that allow them to truly showcase their worth; tests, in other words, like CTS 23–5262.[416] At bottom, then, if CTS is to keep pursuing profits (and firearms examiners will not vary their requests from their norm), then the courts must make themselves felt as consumers, must demand, by using the tools of admissibility and exclusion, the rigorous testing that has been denied them for the more than a century[417] that firearms examination evidence has run wild through the criminal legal system.[418] They must nail the coffin shut on the field of firearms examination and leave it to practitioners to break out and back into the courts (with more and better data) if ever they can.

414. *See* PCAST Report, *supra* note 3, at 106 (Indeed, even the impetus for that first, pre-PCAST sample-to-sample study came from outside the field).

415. Mark Twain, *The Man That Corrupted Hadleyburg*, *in* The Man That Corrupted Hadleyburg and Other Stories and Essays 46 (Harper & Brothers 1900).

416. This analysis assumes that firearms examiners will continue to perform comparisons using a comparison microscope and their subjective judgment. But judicial pressure might also spark greater research into, and adoption of, more objective methods. *See e.g.*, PCAST Report, *supra* note 3, at 113 ("One direction is to continue to improve firearms analysis as a subjective method . . . A second—and more important—direction is (as with latent print analysis) to convert firearms analysis from a subjective method to an objective method"); Cuellar et al., *supra* note 30, at 13 (discussing the benefits of objective, automated methods).

417. James E. Hamby, *The History of Firearm and Toolmark Identification*, 31 AFTE J. 26 (1999).

418. Of course, by focusing on courts and admissibility decisions, this Article does not suggest that other, regulatory approaches might not also bear fruit. However, a thorough analysis of such approaches goes beyond its scope. *See e.g.*, Garrett & Mitchell, *supra* note 43, at 954–57.