South Carolina Bar
Continuing Legal Education Division

2026 SC BAR CONVENTION

What Litigators Need to Know about the 'Science' Behind Police-Use-of Force Expertise: Force Science under Daubert

Saturday, January 24

SC Supreme Court Commission on CLE Course No. 260148

South Carolina Bar

Continuing Legal Education Division

Geoffrey Alpert
Ian Adams
&
Seth Stoughton

Original Article

Forced Science: A Critical Appraisal of the Scientific Rigor of 'Force Science' Policing Research

Police Quarterly
2025, Vol. 0(0) 1–39

DOI: 10.1177/10986111251357498
journals.sagepub.com/home/pqx


Ian T. Adams[1], Seth Stoughton[1], Brandon del Pozo[2], Irick T. J. Geary[1], Marc Olson[1], and Geoffrey P. Alpert[1]

Abstract

Courts regularly rely on scientific expert testimony in police use-of-force cases, and the Force Science Institute (FSI)—a private, for-profit entity—has emerged as an important source of purportedly scientific principles shaping police training, policies, and legal outcomes. Its "peer-reviewed" corpus is presented as scientifically authoritative and feeds into "Force Science Analyst" certification and expert consulting. Employing bibliometric analysis and three validated measures of scientific reliability, we assess a recent volume (*Force Science*) representing FSI's claimed knowledge for scientific reliability. Contrary to FSI's assertions, our findings show its published materials fail to meet the scientific rigor demanded by the *Daubert* standard, which governs the admissibility of scientific evidence in U.S. courts. These results highlight the need for caution and critical scrutiny of such evidence, and suggest that reliance on *Force Science* in legal proceedings, training programs, and policing policies risks introducing unverified concepts into high-stakes decision-making contexts.

Keywords

testimony, force science, reliability, police, use of force, *Daubert*, *Kumho Tire*, *Frye*

[1]University of South Carolina, Columbia, SC, USA
[2]Brown University, Providence, RI, USA

Corresponding Author:
Ian T. Adams, University of South Carolina, Currell Building #114, Columbia, SC 29208-0001, USA.
Email: ian.adams@sc.edu

Introduction

Since the 1908 "Brandeis Brief" (Muller v. Oregon, 1908), US courts have relied upon social science findings to adjudicate issues across a wide spectrum of cases, including "discrimination, the death penalty, and placing a monetary value on a life" (Smith & Alpert, 2002, p. 687). The intersection of scientific evidence and legal decision-making represents a critical frontier in criminal justice, particularly concerning police use-of-force incidents. Courts' increasing reliance on scientific expertise to inform and assist the fact-finder, ultimately influencing judicial outcomes, stresses the need for rigorous research on police behavior that adheres to well-established scientific standards (Fournier, 2016; Reisberg et al., 2016).

Within this context, the Force Science Institute (FSI), a private, for-profit firm, has promoted itself as the foremost source of scientific evidence relevant to use-of-force litigation, amplifying its influence through the certification of "Force Science Analysts" and consulting on high-profile law enforcement cases (Apuzzo, 2015; Nave et al., 2024; Valentino-DeVries et al., 2021). FSI's concepts shape the broader discourse on use-of-force practices, informing police training and policy development, investigations conducted by police oversight entities, and public debate.

FSI's prominent role in shaping police policy and influencing legal outcomes has attracted mounting attention from journalists (Apuzzo, 2015), as well as researchers and other experts. Each of these groups has raised questions about the scientific validity of FSI's methodological approaches and the conclusions in FSI's publications, offering critical assessments in peer-reviewed publications (Fournier, 2016; Hyman, 2022; Nave et al., 2024), academic commentary (Hyman, 2022), and expert testimony (Fournier, 2011, 2019).

Recent scholarship has critically examined specific courtroom uses of FSI "expertise," highlighting how claims regarding officers' perception-reaction times and post-incident memory reliability are leveraged in use-of-force litigation. Nave et al. (2024) provide a careful rendering of several cases where FSI-affiliated experts asserted scientifically-framed claims—such as the inevitability of delayed reaction times that justify officers shooting preemptively and the necessity of two sleep cycles before interviewing involved officers—in ways that have directly influenced jury interpretations of "officer reasonableness." However, Nave and colleagues illustrate that these claims often involve significant methodological shortcomings, raising concerns about their scientific validity. For example, the authors recount the prosecution's expert witness in a murder trial involving a police officer defendant, characterizing FSI research as "…a very clinical situation. It was very, very clean. It was done in a laboratory-type atmosphere" (Nave et al., 2024, p. 15). Despite the appeal to scientific authority and language, the underlying material relied upon by the expert here was, in fact, not scientifically reviewed or published, and instead appeared in the non-scientific, professional publication *Law Enforcement Executive Forum* (Lewinski et al., 2014).

Given these concerns about FSI's role as a purveyor of putatively *scientific* information, concerns succinctly articulated in Nave et al. (2024), FSI's publications

would benefit from the standard types of review and scrutiny accorded to research presented as credible scientific evidence. There are two compelling reasons for such scrutiny. First, the integrity of scientific evidence introduced in court relies on robust, generalizable, and replicable research, as encapsulated by the *Daubert* standards of testability, peer review, error rate, and general acceptance. An examination of FSI publications would inform courts and other legal decision makers in the application of the *Daubert* standard. Second, evidence-based policing relies on high-quality empirical research to inform policy and practice, requiring accountability and scientific rigor from its sources to meet this commitment. Without scrutiny, the scientific quality of research by any source, including FSI, remains a matter of conjecture and speculation. Yet no comprehensive evaluation of FSI's research corpus exists.

Our goal, therefore, is to clarify the extent to which of FSI's contributions should inform legal proceedings and evidence-based policing practices by assessing whether FSI's publications are consistent with widely acknowledged standards of rigor within the scientific community. In doing so, we provide a general framework for prospectively assessing scientific research that can influence critical decisions in policing and assist in the development of valid empirical evidence for use in police practice and the courtroom.

Scientific Evidence & The Force Science Institute

The legitimacy of legal proceedings depends on fair and accurate adjudication, which requires litigants to substantiate their claims with credible evidence (Walrdon, 2011). Evidence law—formalized, *inter alia*, through the Federal Rules of Evidence—provides a framework for identifying admissible evidence and specifying how it may be used in judicial proceedings. Within this system, expert testimony plays a pivotal role, providing specialized knowledge about topics beyond the common understanding of factfinders (Federal Rule of Evidence 702). A longstanding challenge for courts, however, is distinguishing credible expert assertions from claims lacking in empirical or methodological support (Chan, 1995; Huber, 1993).

The Paradoxical Nature of Scientific Evidence

In our legal system, jurors must determine how much weight they will give to various pieces of evidence. For example, jurors may assess that one witness is more credible than another or that a video refutes a witness's testimony. Outside of the investigative grand jury process, however, jurors do not determine what evidence to review. Instead, lawyers seek to introduce evidence, and judges determine whether that evidence is admissible. Expert testimony is admissible when based on sufficient "scientific, technical, or other specialized knowledge." (Federal Rule of Evidence 702)

Admissibility is particularly important in the context of scientific expertise. Presenting expertise in scientific terms has the distinct capacity to serve as what scholars term a "veneer" – a way to make conclusions more palatable and convincing even when

they "conflict with mundane reasoning and challenge credibility" (Nave et al., 2024, p. 1). As Nave et al. (2024, p. 9) observe, use-of-force experts' "scientific" claims carry particular weight because "court procedures problematize the intersubjective world in ways that set the stage for rendering an objective solution… and scientific claims fit easily within that frame." This influence stems not only from procedural dynamics but also from (p. 9) "a long history of scientific and social scientific expertise yielding insightful answers to questions put before the court."

The perceived objectivity of the scientific method amplifies this dynamic. Organizations such as FSI explicitly leverage this perception, promoting their work as "the research and application of unbiased scientific principles and processes to determine the true nature of human behavior in high stress and deadly force encounters" (Force Science, 2018a). This framing strategically positions FSI claims as inherently credible and impartial *because they are scientific*, enhancing their rhetorical force in legal proceedings (Nave et al., 2024).

Herein lies the paradox. While judges and jurors typically understand that the scientific method is used to generate reliable knowledge, they often lack the specialized expertise necessary to distinguish between theories or techniques that reflect rigorous scientific inquiry and those that convincingly employ scientific language as a veneer. In short, there is often a substantial gap between an expert's assertions of scientific authority and factfinders' ability to understand and critically assess the basis for that authority.

To address this vulnerability, courts predicate the admissibility of scientific evidence on a showing that such evidence meets a minimum threshold of reliability. As the Advisory Committee note to Rule 702 emphasizes, "[j]udicial gatekeeping is essential because just as jurors may be unable, due to lack of specialized knowledge, to evaluate meaningfully the reliability of scientific and other methods underlying expert opinion, jurors may also lack the specialized knowledge to determine whether the conclusions of an expert go beyond what the expert's basis and methodology may reliably support" (Committee Notes on Rules—2023 Amendment, 2023).

The Admissibility of Scientific Evidence

For years, the admissibility of scientific evidence was governed by the standard set by the D.C. Court of Appeals in *Frye v. United States*,[1] which specified that expert testimony was admissible so long as the underlying methodology was "generally accepted" within the relevant scientific field. Despite widespread adoption by federal and state courts, *Frye* was criticized for permitting the admission of evidence derived from methods that, while generally accepted, lacked scientific rigor and reliability, and for excluding evidence that resulted from the application of novel. but empirically sound, scientific methods. In 1975, Congress adopted the Federal Rules of Evidence, including Rule 702, which revised the criteria for admitting scientific evidence by requiring that expert testimony be based on reliable principles and methods (Giannelli, 1980).

In 1993, *Daubert v. Merrell Dow Pharmaceuticals*[2] further refined Rule 702, holding that scientific testimony must be "scientifically valid" (p. 593) and "reliable" (p. 589). The Court specifically noted that it was not adopting the *scientific* definitions of validity and reliability, which narrowly pertain to study design and replication, but rather used those terms to describe "evidentiary reliability," or the overall "trustworthiness" of the evidence (p. 590, fn. 9). In short, to meet the standard of evidentiary reliability, scientific evidence—including expert testimony—must be scientifically credible (i.e., both valid and reliable).[3]

The Court did not delve into the precise practices and standards that establish scientific credibility, but it did establish guidelines to enhance the scientific rigor of expert testimony presented in court, aligning with broader scientific standards of accuracy and objectivity (Chan, 1995; Faigman, 2012; Faigman et al., 1999; L. R. Fournier, 2016; Gatowski et al., 2001). The *Daubert* Court laid out four factors for courts to assess the trustworthiness of scientific evidence: *testability, peer review and publication, error rates, and general acceptance by the scientific community.*

- *Testability* addresses whether an expert's claim, theory, or technique has been—or can be—empirically tested, a foundational element of scientific inquiry.
- *Peer review and publication* refers to whether the evidence underpinning the expert's testimony is supported by peer-reviewed studies that have undergone scrutiny for internal validity and statistical reliability.
- *Error rate* pertains to the known or potential frequency of error in the scientific method or technique behind the expert's claim, addressing internal, statistical, and external validity relevant to the specific case.
- *General acceptance* evaluates the degree of consensus among experts in the relevant scientific field regarding the methods underlying the expert's testimony.

These four *Daubert* criteria were further entrenched by later amendments to Rule 702, providing courts with a structured framework to evaluate the credibility of scientific evidence. Citing Daubert dicta, Chan (1995, p. 11) notes that *Daubert* is aimed at ensuring scientific validity, which is "not so much a standard or a benchmark as it is an inquiry into *process*, centered 'solely on principles and methodology, not on the conclusions they generate.'"

Rule 702 and the Daubert framework help courts identify and exclude "junk science," thereby ensuring that only scientifically rigorous evidence informs judicial decisions (Huber, 1993). These four considerations—testability, peer review, error rates, and general acceptance—are essential to establishing the scientific credibility upon which evidentiary reliability rests. Requiring scientific expert testimony to adhere to stringent standards serves an important jurisprudential function, preserving judicial accuracy and fairness (Fournier, 2016).[4]

Internal and External Validity

The *Daubert* standard helps ensure that scientific evidence satisfies two overarching benchmarks: internal validity and external validity. *Internal validity* exists when research results are accurate, i.e., that a study measures what it claims to measure. This can be ensured, for example, through random assignment of treatment and controls and conducting studies with fidelity to scientific principles. *External validity* exists when research results can be generalized, or reliably applied to circumstances and settings beyond those specifically tested in the underlying research. This can be achieved, for example, through large or representative sampling.

Scientific expert testimony requires the underlying research to be both internally and externally valid. This observation cannot be overemphasized: In *every instance* where an expert witness offers scientific testimony, they are applying prior research results to the facts at hand. When the underlying research suffers from poor design or implementation, we cannot be confident that its conclusions are valid. And when we cannot be confident that the conclusions are valid in the context of the prior research (*internal*), we cannot confidently apply those conclusions to a novel situation, such as the one being contested at trial (*external*). Thus, internal validity is a necessary, but not sufficient, condition for external validity. Moreover, even when a study demonstrates internal validity, it is of no use in litigation unless its findings can be reliably generalized to the case at bar.

FSI Training and Concepts

The Force Science Institute (FSI) exerts considerable influence over how practitioners and judicial actors evaluate police use-of-force incidents, primarily through its extensive, widely attended training programs. As Nave et al. (2024, p. 14) inform us, "[T]he FSI has become a clearinghouse in the expertise market, an example of what some scholars describe more generally as reflective of 'the commodification of the expert.'"

FSI Training and Consulting. FSI's educational curricula comprises three distinct courses focused on use-of-force incident evaluation. Their entry-level "Force Encounters Course" consists of 16 hours of instruction ($395) covering "psychological and physiological factors that affect threat assessment, sensory perception, decision, performance, and memory and applies these evidence-based concepts to high-stress and life-threatening encounters in a law enforcement context" (Force Science, 2022a). The institute emphasizes that this curriculum derives from "peer reviewed research that uses precise time-and-motion measurements to document environmental, physiological, and psychological dynamics during high-threat events," claiming their approach "support[s] commitment to procedurally just investigations, employing realistic, thorough, and evidence-based analysis" (Force Science, 2022a).

The institute's core certification program, which produces hundreds of "Force Science Analysts" annually, operates through both hybrid ($1,650) and online-only

($1,399) formats (Force Science, 2019). According to FSI, these programs "use … the combined knowledge of a team of world-class experts to explore the complex mix of human dynamics involved in the often complicated, confusing, and controversial uses of force by law enforcement personnel" (Force Science, 2022b). The stated objective is to enhance law enforcement professionals' understanding of human factor dynamics, allowing "graduates to draw clear, accurate, and concise conclusions in their [use-of-force] investigations" (Force Science, 2022b).

FSI's advanced offering, the "Advanced Force Science Specialist Course" (price available only upon request), purports to examine "the intricate facets of human behavior during critical incidents, anchored in solid research and theory about high-stress human performance" (Force Science, 2023). The curriculum claims to span multiple disciplines, including "fundamental motor actions, their role during pivotal moments, attention dynamics like eye tracking and focus, pattern discernment, proactive thinking, decision-making processes, and memory recall" while incorporating elements of "neurophysiology, vision, cognition, memory, decision-making, learning theory, exercise physiology, and kinesiology" (Force Science, 2023).

Beyond its educational offerings, FSI maintains a Consulting Division frequently involved in officer legal defense during controversial cases. The division claims its members possess unique qualifications in explaining "human performance during dynamic force encounters" and "specialize in understanding the psychological, physiological, and environmental influences that can impact perception, attention, and performance during police and civilian-involved critical incidents" (Force Science, 2018b).

Influence of FSI Training and Concepts on Legal Processes. As articulated by its founder, FSI actively seeks to shape understanding across the criminal justice landscape, aiming to "educate law enforcement, courts, and communities" about "use-of-force decision-making, performance, and outcomes" (Lewinski, 2022, p. XV). This influence manifests through three distinct but interrelated mechanisms: litigation outcomes, pre-litigation decision-making, and police training.

First, expert testimony by Certified Force Science Analysts can significantly shape factfinder determinations in use-of-force proceedings. High-profile cases illustrate this dynamic. In the 2021 prosecution of former Brooklyn Center, Minnesota officer Kim Potter for the fatal shooting of Duante Wright, defense experts employed FSI's "slip and capture" concept, introducing it as a type of "action error," to explain how Potter mistook her pistol for her Taser when she fired the fatal shot. Similarly, FSI founder Dr William Lewinski previously testified about "slip and capture" and "inattentional blindness" in defense of Bay Area Rapid Transit officer Johannes Mehserle, who was being prosecuted for the fatal shooting of Oscar Grant III. And in the 2024 trial of Connecticut State Trooper Brian North, prosecutors initially retained Sgt. Jamie Borden, an FSI Senior Instructor, to analyze a police use-of-force incident. After Borden recommended against charging the officer involved, the defense subsequently engaged him to support their case (Brown, 2024). In all three cases, FSI-derived concepts lent scientific legitimacy to defense arguments challenging perception or

intentionality, with tangible legal consequences—as evidenced by Mehserle's conviction for manslaughter but acquittal for murder (Lewinski, 2010; Remsberg, 2011).

Second, FSI concepts systematically influence pre-judicial decision-making by civil and criminal attorneys. FSI actively cultivates this influence through its Consulting Division, explicitly targeting prosecutors and district attorneys in its case review intake process; when attorneys request a case review on the FSI website (Force Science, 2024), the second and third question on the form are, respectively, "Are you a prosecutor?" and "Are you a district attorney?" The impact is concrete: the Plymouth County, Massachusetts's District Attorney's Office, for instance, cited the evaluation provided by a "Certified Force Science Analyst" in declaring a trooper's use of lethal force "appropriate and lawful" (Plymouth County District Attorney's Office, 2021, p. 4). In January 2023, the California Attorney General's office cleared Los Angeles Police Department Officer Toni McBride of wrongdoing in the 2020 fatal shooting of Daniel Hernandez, relying in part on the analysis of Dr Lewinski, whom the office hired to review the case (Rector, 2023). This sequence illustrates how FSI analyses – rooted in *Force Science* studies – can influence prosecutorial decision-making on whether and how to criminally charge officers.

Third, FSI concepts shape the institutional processes that determine what information reaches legal decision-makers. As noted on the website Police1, agency administrators are encouraged to review tactics "through the eyes of a Force Science-certified analyst" (Blake, 2013). Certain FSI principles, such as "action beats reaction," have achieved what Nave et al. (2024, p. 12) describe as the status of "common knowledge held among police officers," fundamentally framing how use-of-force incidents are evaluated and articulated. This, as with other FSI concepts, has obvious implications for internal evaluations of use-of-force incidents, as well as how those internal evaluations are communicated to external audiences that precede legal decision-making. In British Columbia, Canada, for example, a Force Science Analyst explained to a civilian oversight board why "pre-assaultive cues"—"'squaring off in a combat stance,' clench[ed] hands, exhibiting a '1000-yard stare,' and displaying 'resistive tension'"—justified an officer's use of "pre-emptive force" against a motorist, doing so by "'fram[ing] the 'cues' issue in the context of 'a concept spelled out in scientific detail'" (Remsberg, 2011, para. 36).

The penetration of FSI concepts into police-related litigation exemplifies what Lvovsky (2017, p. 2081) termed the "judicial presumption of police expertise." The pervasive influence of FSI concepts across legal and administrative domains is not incidental but strategic, as evidenced by Force Science News headlines celebrating their impact: "Case Studies: How Force Science Analysts Helped Accused Officers" (Remsberg, 2011); "Top Experts Work with Force Science to Advance Police-Related Research" (Kliem, 2022); "Trainers as Police Practice and Human Factors Experts" (Kliem, 2023); "Leading the National Discussion on Policing" (Kliem, 2020); "Aligning Research on Human Performance Across High-Stakes Professions" (Kliem, 2024).

Methods

Given FSI's influence on law enforcement training and policy development, its frequent participation in use-of-force litigation (Nave et al., 2024), and the stated goals of FSI itself to influence courts, policy, and police training (Lewinski, 2022), the reliability of its research should be evaluated against relevant evidentiary standards.

To do so, this study undertakes a systematic appraisal of the FSI research reproduced in a recent volume entitled *Force Science: Peer-Reviewed Scientific Research* (hereafter, *Force Science*). The volume consists of twenty-four previously published articles that Dr Lewinski, FSI's founder, director, and senior researcher, describes as addressing "some of the most critical questions asked by the Force Science research team," documenting findings that have substantial implications for "investigations, training, and honest accountability" in law enforcement (Lewinski, 2022, p. XV). The centrality of these articles to FSI's mission is evident in the explicit aim of the collected volume: to employ scientific principles and evidence to explain the complex interactions between human and environmental factors in police use-of-force scenarios, with the stated purpose of educating "law enforcement, courts, and communities" (Lewinski, 2022, p. XV). Thus, *Force Science* represents the core scientific bases of FSI's curricula, training, reports, testimony, and research agenda. These articles are also reflective of FSI's research methods, as evidenced by Dr Lewinski's first authorship of ten of the studies and co-authorship of the remaining fourteen.

The appraisal here employs established assessments of scientific reliability and impact, examining the methodological practices, transparency, and fulfillment of the *Daubert* criteria of the studies contained in *Force Science*. The analysis then considers how these standards shape the suitability of FSI research for judicial and policy applications. This evaluation is particularly timely given the growing national emphasis on evidence-based policing and the need for science-driven approaches to police reform (Lum & Koper, 2017; Ratcliffe, 2022). Our study provides insights with broad implications for its suitability as a source of courtroom evidence and for the policing policies that rely on FSI for their justification.

As an analysis of secondary data in the public domain, it was exempt from IRB review as per the Common Rule.

Data

Our analysis focuses on the twenty-four articles compiled in *Force Science,* a collection that the FSI characterizes as its core scientific contributions and that serves as both the scientific foundation for FSI's training programs and the principal scientific bases for their investigative methodologies. *Force Science* groups these articles into eight thematic areas: Foundational Principles, Attention and Vision, Memory, Speed and Dynamics, Forensic Considerations, Human Error, Equipment, and Training.

Analytical Plan

Our analysis consists of two components: (1) An assessment of the bibliometric impact of the articles collected in *Force Science*, and (2) an analysis of their scientific rigor.

Bibliometric assessment of impact. To assess the *Force Science* articles' acceptance by, and impact on, the scientific community, we conduct a bibliometric analysis. A key consideration for evaluating scientific evidence offered in a courtroom, and for academic evaluation of scholarly work, is the extent to which the relevant scientific discipline relies upon those findings. Rough measures of this reliance can be traced by identifying (1) the citations to a given article, and (2) the journal impact factor for the scientific outlet where the article was published. These indicators, as general proxies for research quality and reach, provide standardized measures of scholarly influence and peer recognition within the scientific community, and they are commonly used to evaluate the scholarly work of scientific researchers.

To assess scientific impact, we employ four complementary metrics: Web of Science citations, Category Normalized Citation Impact (CNCI), Journal Normalized Citation Impact (JNCI), and Journal Impact Factors (JIF), all measured as of November 18, 2024. These metrics provide valuable insight into scholarly engagement and influence. Web of Science citations reflect direct academic engagement, indicating how often peers reference the work. CNCI) offers a discipline-normalized measure of citation performance relative to articles within the same research category, allowing for comparisons across different fields. JNCI similarly contextualizes citation impact relative to the publishing journal's typical performance. Lastly, JIF represents the overall scholarly reputation of the journal, suggesting the expected visibility and credibility of articles published therein.

Collectively, these metrics assess whether the Force Science Institute's (FSI) research portfolio is well-accepted and influential within the broader scientific community. These metrics offer distinct perspectives on scholarly influence, so while no single metric definitively establishes scholarly impact (DeJong & St George, 2018), the triangulation of these measures provides insight into the scientific community's engagement with FSI research, and are commonly used across academia to assess the quality of both individual papers as well as scholarly achievement.

Evaluation of rigor. To assess scientific rigor, we used three established methodological assessment tools chosen for their complementary strengths in evaluating diverse research designs. While numerous critical appraisal tools exist, we selected three widely accepted tools with specific relevance to the FSI research corpus, balancing breadth and relevance, prioritizing tools that assess causal inference, study design quality, and adaptability across quantitative, qualitative, and mixed-methods studies. These tools are:

- The Maryland Scientific Method Scale (Farrington et al., 2002; Madaleno & Waights, 2016) – chosen for its assessment of causal inference capacity in studies impacting public policy.
- The Newcastle-Ottawa Quality Assessment Scale (adapted for cross-sectional studies) (Modesti et al., 2016) – focuses on study group selection, comparability, and outcome/exposure ascertainment, key for non-randomized research.
- The Mixed Methods Appraisal Tool (Hong, Gonzalez-Reyes et al., 2018; Hong, Pluye, et al., 2018) – provides a flexible framework suited to diverse research designs, including qualitative, quantitative, and mixed methods.

This multi-tooled approach provides for a nuanced evaluation of methodological quality in order to appraise the heterogeneous research approaches in FSI's body of work. Details about each of the three appraisal tools are included in the appendix.

Coding Methodology

The coding team comprised three individuals: the lead author, who holds both basic and advanced Force Science certification and has professional policing experience; a graduate student with basic Force Science certification and extensive professional experience; and another graduate student with limited civilian analyst experience in a police agency but no direct Force Science training. This composition was intentionally designed to incorporate different perspectives and substantive experience, including those without direct Force Science backgrounds.

Coders participated in two training sessions totaling 4 hours. The initial session provided an overview of the *Force Science* article collection and introduced the three scoring systems. The subsequent session delved into the specifics of each scoring system's components, complemented by practical exercises where coders collectively applied the scoring systems to a specific article until they demonstrated a consistent approach.

Coders conducted their evaluations independently, applying the predefined scoring benchmarks to each article to minimize bias and ensure a fair evaluation based on established criteria. Following the independent evaluations, the research team gathered to review the findings collectively. This phase was crucial for evaluating difficult-to-code cases, identifying consistencies and inconsistencies among the assessments, and discussing the reasons behind these findings. This collaborative review process served as an additional layer of quality control, enhancing the reliability and validity of the evaluation outcomes. The structured discussion and example-based training led to a consistent, coherent, and robust understanding of evaluative metrics among reviewers.

In their assessments, coders worked from a prepared template listing the 24 articles collected in *Force Science*. Importantly, coders evaluated the original versions of these articles as they were, rather than the versions compiled in the book. This approach was chosen to ensure the study's findings would be directly relevant to legal contexts, recognizing that the original publications are the versions most likely to be cited in

Table 1. Scoring Scale Inter-rater Reliability.

Measure	Range	Raw Agreement %	IRR	Lower CI	Upper CI	*p*
Maryland Scientific Methods Scale (MSMS)						
Maryland Implementation	0–5	90.91	0.94	0.89	0.97	0.00
Maryland Method	0–5	90.91	0.98	0.95	0.99	0.00
Mixed Methods Appraisal Tool (MMAT)						
MMAT type	1–5	90.91	0.94	0.88	0.97	0.00
MMAT 1	0–1	100.00	1.00	n/a	n/a	0.00
MMAT 2	0–1	81.82	0.75	0.57	0.88	0.00
MMAT 3	0–1	95.45	0.86	0.75	0.94	0.00
MMAT 4	0–1	81.82	0.76	0.58	0.88	0.00
MMAT 5	0–1	86.36	0.82	0.68	0.91	0.00
Newcastle-Ottawa Quality Assessment (NOQ)						
Selection	0–5	87.50	0.88	0.79	0.94	0.00
Comparability	0–2	81.82	0.81	0.66	0.91	0.00
Outcome sum	0–3	91.67	0.84	0.71	0.92	0.00

Note. MMAT Type, Maryland Method, and Implementation report the Fleiss' Kappa. All other measures report Intraclass Correlation (ICC). ICC requires variability in the ratings to assess the correlation effectively. Items with full agreement do not calculate an IRR/Kappa.

court documents and by expert witnesses, especially considering the book's publication in 2024.

Two studies (IDs 1 and 6) were excluded from scoring due to their status as narrative reviews rather than empirical research. Consequently, no quantitative scores were assigned, as the assessment tools used in this review are designed for evaluating empirical studies.

The reliability of the coders' assessments was evaluated using inter-rater reliability (IRR) measures. The results, detailed in Table 1, demonstrate high levels of agreement across the scoring systems, indicating consistent and reliable evaluations. The Maryland Scientific Methods Scale (MSMS) showed an agreement percentage of 90.91% with high IRR for both implementation (0.94) and method (0.98) assessments. The Mixed Methods Appraisal Tool (MMAT) exhibited varying levels of agreement, with perfect agreement (100%) for MMAT 1 and high agreement for other components (ranging from 81.8% to 95.5%), reflecting an IRR between 0.75 and 0.94. The Newcastle-Ottawa Quality Assessment (NOQ) also demonstrated high reliability, with agreement percentages ranging from 87.5% to 91.7% and IRR values between 0.81 and 0.88. These results highlight the robustness of the scoring systems and the reliability of the coders' evaluations, as concordance is within the thresholds commonly cited in the literature on interrater reliability (Hallgren, 2012; Landis & Koch, 1977).

Table 2. Article Corpus Under Review.

ID	Title	Journal	Year	Section	Web of Science Citations	Category Normalized Citation Impact[a]	Journal Normalized Citation Impact[b]	Journal Impact Factor (2023)
1	A survey of the research on human factors related to lethal force encounters: Implications for law enforcement training, tactics, and testimony	Law Enforcement Executive Forum	2008	Foundational Principles	0	-	-	0
2	The attention study: A study on the presence of selective attention in firearms officers	Law Enforcement Executive Forum	2008	Attention and Vision	0	-	-	0
3	Force science forum: Command types used in police encounters	Law Enforcement Executive Forum	2008	Attention and Vision	0	-	-	0
4	Command sequence in police encounters: Searching for a linguistic fingerprint	Law Enforcement Executive Forum	2008	Attention and Vision	0	-	-	0
5	Pursuit driver training improves memory for skill-based information	Police Quarterly	2008	Memory	1	Below Average	Below Average	2.9
6	New developments in understanding the behavioral science factors in the "stop shooting" response	Law Enforcement Executive Forum	2009	Speed and Dynamics	0	-	-	0
7	Fired cartridge case ejection patterns from semi-automatic firearms	Investigative Sciences Journal	2010	Forensic Considerations	0	-	-	0

(continued)

Table 2. (continued)

ID	Title	Journal	Year	Section	Web of Science Citations	Category Normalized Citation Impact[a]	Journal Normalized Citation Impact[b]	Journal Impact Factor (2023)
8	Performing under pressure: Gaze control, decision making and shooting performance of elite and rookie police officers	Human Movement Science	2012	Attention and Vision	90	Above Average	Above Average	1.6
9	Witnesses in action: The effect of physical exertion on recall and recognition	Psychological Science	2012	Memory	36	Above Average	Below Average	4.8
10	The influence of officer positioning on movement during a threatening traffic stop scenario	Law Enforcement Executive Forum	2013	Speed and Dynamics	0	-	-	0
11	The influence of start position, initial step type, and usage of a focal point on sprinting performance	International Journal of Exercise Science	2013	Speed and Dynamics	0	-	-	0
12	Police officer reaction time to start and stop shooting: The influence of decision-making and pattern recognition	Law Enforcement Executive Forum	2014	Speed and Dynamics	0	-	-	0
13	Police officers' actual versus recalled path of travel in response to a threatening traffic stop scenario	Police Practice and Research	2016	Memory	13	*n/a published prior to JIF awarded	*n/a published prior to JIF awarded	1.4

(continued)

Table 2. (continued)

ID	Title	Journal	Year	Section	Web of Science Citations	Category Normalized Citation Impact[a]	Journal Normalized Citation Impact[b]	Journal Impact Factor (2023)
14	The real risks during deadly police shootouts: Accuracy of the naïve shooter	International Journal of Police Science & Management	2015	Speed and Dynamics	0	-	-	0
15	Ambushes leading cause of officer fatalities-- when every second counts: Analysis of officer movement from trained ready tactical positions	Law Enforcement Executive Forum	2015	Speed and Dynamics	0	-	-	0
16	The influence of officer equipment and protection on short sprinting performances	Applied Ergonomics	2015	Speed and Dynamics	25	Below Average	Above Average	3.1
17	The speed of a prone subject	Law Enforcement Executive Forum	2016	Speed and Dynamics	0	-	-	0
18	Memory and the operational witness: Police officer recall of firearms encounters as a function of active response role	Law and Human Behavior	2016	Memory	38	Above Average	Above Average	2.4
19	Toward a taxonomy of the unintentional discharge of firearms in law enforcement	Applied Ergonomics	2017	Human Error	10	Below Average	Below Average	3.1

(continued)

Table 2. (continued)

ID	Title	Journal	Year	Section	Web of Science Citations	Category Normalized Citation Impact[a]	Journal Normalized Citation Impact[b]	Journal Impact Factor (2023)
20	Law enforcement memory of stressful events: Recall accuracy as a function of detail type	Law Enforcement Executive Forum	2017	Memory	0	-	-	0
21	Protective vests in law enforcement: A pilot survey of public perceptions	Journal of Police and Criminal Psychology	2017	Equipment	0	*n/a published prior to JIF awarded	*n/a published prior to JIF awarded	1.6
22	Further analysis of the unintentional discharge of firearms in law enforcement	Applied Ergonomics	2018	Human Error	10	Below Average	Below Average	3.1
23	Training and safety: Potentially lethal blue-on-blue encounters	Police Practice and Research	2019	Human Error	2	*n/a published prior to JIF awarded	*n/a published prior to JIF awarded	1.8
24	Police academy training, performance, and learning	Behavior Analysis in Practice	2019	Training	21	*n/a published prior to JIF awarded	*n/a published prior to JIF awarded	2.1

[a]Category Normalized Citation Impact (CNCI) is the ratio of a document's actual times cited count to the expected count for a document of the same type, from the same category, and published in the same year. If the ratio is above 1, then the document's citation performance is above average.
[b]Journal Normalized Citation Impact (JNCI) is the ratio of a document's actual times cited count to the expected count for a document of the same type, from the same journal, and published in the same year. If the ratio is above 1, then the document's citation performance is above average.

Results

Citation Impact and Reliance

Table 2 details each article's bibliographic information and impact metrics. For all 24 articles in *Force Science*, we document title, journal, publication year, book section placement, and four key impact measures: Web of Science citations, Category Normalized Citation Impact (CNCI), Journal Normalized Citation Impact (JNCI), and journal impact factors (JIF). The data reveals a striking overall lack of scholarly influence.

Of the twenty-four studies collected in *Force Science,* fourteen (58.3%) do not appear in journals recognized by the Web of Science. Seven articles received a CNCI score, with three receiving below average scores. Of the seven articles that received a JNCI score, four received a below average score. Articles published in mainstream scientific journals such as *Psychological Science* and *Applied Ergonomics* exhibit notably higher citation counts, CNCI, JNCI, and JIF values, reflecting stronger engagement and relevance within their respective scientific communities.

In contrast, publications appearing in specialized venues, particularly the *Law Enforcement Executive Forum* (LEEF), demonstrate consistently negligible impact, with CNCI and JNCI scores absent entirely due to their exclusion from recognized scientific indexing. Ten articles (41.7%) published in LEEF exhibit zero citations in Web of Science, CNCI, JNCI, or JIF metrics. This low impact likely stems from the fact that LEEF is not recognized as a scientific journal; its publisher, the Illinois Law Enforcement Training and Standards Board Executive Institute, does not position itself as a scientific entity (Illinois Law Enforcement Training and Standards Board Executive Institute, 2024). Further, scholars specifically note that LEEF lacks the peer-review process essential for ensuring scientific validity and reliability (Fournier, 2011). William Lewinski, the sole common author of all twenty-four *Force Science* articles, is listed as serving as an Associate Editor of *LEEF* (Illinois Law Enforcement Training and Standards Board Executive Institute, 2015).

The reporting in Table 2 shows the limited scientific reach of the Force Science Institute (FSI) portfolio. Aggregated metrics reflect modest citation levels across all twenty-four articles: Web of Science citations (M = 10.25, SD = 20.6), Journal Impact Factor (M = 1.16, SD = 1.45), and frequent absences of meaningful CNCI and JNCI scores. Notably, the substantial gap between mean and median values highlights highly skewed distributions, with median values (Web of Science = 0, JIF = 0) significantly below the mean, emphasizing that the typical FSI publication garners minimal scholarly impact. The mode for all primary indicators (Web of Science, CNCI, JNCI, and JIF) remains zero, further demonstrating the limited scholarly influence of the majority of these publications.

These impact metrics demonstrate significant variation across *Force Science* book sections, as reflected in Table 3, though interpretative caution is warranted for sections containing single articles (Foundational Principles, Forensic Considerations, Equipment, and Training).

Table 3. Descriptive Statistics for Scientific Impact.

Section	n	Metric	Mean	SD	Median	Mode	Range
All Articles	24	Web of Science Citations	10.25	20.6	0	0	0–90
		Journal Impact Factor	1.16	1.45	0	0	0–4.8
Foundational Principles	2	Web of Science Citations	0	-	0	0	0–0
		Journal Impact Factor	0	-	0	0	0–0
Attention and Vision	4	Web of Science Citations	22.5	38.97	0	0	0–90
		Journal Impact Factor	0.4	0.69	0	0	0–1.6
Memory	5	Web of Science Citations	17.6	16.5	13	0	0–38
		Journal Impact Factor	3.12	2.68	2.5	0	0–8.2
Speed and Dynamics	7	Web of Science Citations	3.13	8.27	0	0	0–25
		Journal Impact Factor	0.39	1.03	0	0	0–3.1
Forensic Considerations	1	Web of Science Citations	0	-	0	0	0–0
		Journal Impact Factor	0	-	0	0	0–0
Human Error	3	Web of Science Citations	7.33	3.77	10	10	2–10
		Journal Impact Factor	2.67	0.61	3.1	3.1	1.8–3.1
Equipment	1	Web of Science Citations	0	-	0	0	0–0
		Journal Impact Factor	1.6	-	1.6	1.6	1.6–1.6
Training	1	Web of Science Citations	21	-	21	21	21–21
		Journal Impact Factor	2.1	-	2.1	2.1	2.1–2.1

This heterogeneity in scientific impact metrics underscores the necessity of careful methodological assessment. While citation counts and impact factors provide useful contextual information about scholarly recognition, they cannot be relied upon alone (Krampl, 2019; Martín-Martín et al., 2018). They must be interpreted alongside rigorous evaluation of research design, methodological quality, and analytical robustness, which is where we turn our attention next.

Scientific Rigor

The concentration of FSI's work in specialized law enforcement journals, rather than in mainstream scientific publications, raises important questions about the breadth and depth of peer review and scholarly scrutiny these studies have received. To address those questions, this study next analyzed the scientific rigor of the articles collected in *Force Science*. The critical appraisal of the twenty-two scorable articles[5] in *Force Science* using MMAT, MSMS, and NOQAS reveals significant and endemic methodological flaws.

For empirical research to meet the scientific credibility standard required under *Daubert*—particularly when it is intended to inform judicial and law enforcement training and decision-making—it should demonstrate adherence to basic scientific methodology. Such adherence would naturally result in high scores on standardized quality assessments like the MMAT, NOQ, and MSMS.

Interpreting the scores derived from the application of these tools is *not* akin to school grading systems. High ratings are the threshold indication of scientific credibility and reliability, rather than rare signals of extraordinary achievement. Scoring these tools requires assessing different components of research design and implementation, and courts agree. As the Court of Appeals for the Eighth Circuit has noted, "[A]ny step that renders the analysis unreliable renders the expert's testimony inadmissible. This is true whether the step completely changes a reliable methodology or merely misapplies that methodology."[6] For that reason, even small deficiencies in scoring signal significant concerns about the research's validity, methodological rigor, and reliability. Small shortfalls in key domains (e.g., participant selection, measurement, or analysis) can disproportionately undermine the trustworthiness of a study's conclusions (Hong & Pluye, 2019). Thus, what appears to be a mid-level score on any one of the tools actually reflects a substantial deviation from well-established norms of quality scientific research.

The majority of the articles analyzed here suffer from severe issues in design, execution, and reporting. Only a few demonstrate moderate quality research practices. None demonstrate high quality ratings across scales.

Summary Results

MSMS Summary Results. The MSMS evaluates the methodological rigor and implementation quality of research designs, focusing on their capacity to establish causal relationships and relevance for policy implications. The summary results of the appraisal using the Maryland Scientific Methods Scale (MSMS) are reported in Appendix Table A. The overall results from the MSMS assessment of the *Force Science* articles indicate significant methodological weaknesses across most *Force Science* articles.

The mean score for methodological rigor (Method) was 1.55 (SD = 1.44) out of a possible 5, with scores ranging from 0 to 5. Only two studies (ID 9 and 18) achieved the maximum score, indicating an experimental design intent and the ability to establish strong causal inferences. No other studies received a score higher than 2, and seven studies (IDs 3, 4, 14, 15, 17, 19, and 22) received a score of 0, highlighting significant deficiencies in research design and an inability to establish causality.

Implementation quality (Implementation) scored even lower, with a mean of 0.91 (SD = 1.10) out of 5. Scores ranged from 0 to 4, with the highest score being 4, again achieved by the two experimental studies (IDs 9 and 18). No other studies received a score higher than 1, and seven studies (IDs 3, 4, 14, 15, 17, 19, and 22) received a score of 0, indicating poor implementation and significant methodological flaws.

The relatively low variation in scores demonstrates consistently poor methodological standards applied across the studies, highlighting the need for improved research design and execution in order for the *Force Science* studies to achieve the reliability required to inform policy.

NOQ Summary Results. The NOQ evaluates methodological rigor of research design, focusing on selection, comparability, and outcome. The summary of the NOQ scores for each FSI article reviewed is found in Appendix Table B. Overall, the articles are of low quality, with a mean sum NOQ score of 4.32 (SD = 1.12) out of 10. Only one study was of moderate quality (ID 7, with a NOQ score of 6.67). None of the articles had a consistent high classification, underscoring the overall deficiency in methodological rigor among the evaluated studies. Across the articles, several issues consistently contributed to low scores.

First, the articles demonstrate low reliability in *Selection 1* (M = 0.21, SD = 0.39; out of 1), which asks about sample representativeness; in *Selection 2* (M = 0, SE = 0; out of 1), which asks whether sample size is justified and satisfactory; and in *Selection 3* (M = 0, SE = 0; out of 1), which asks whether the study adequately addresses comparability between participants and non-participants. These low scores are driven primarily by two factors: sample size and convenience sampling. The sample sizes are consistently quite small, and in no case did a single article provide a justification for their sample size (e.g., through an *a priori* power analysis). Additionally, the *Force Science* articles largely rely on convenience sampling. Convenience sampling introduces significant bias, as the sample is unlikely to be representative of the target population (in most cases, the population of police officers). The lack of randomization further exacerbates this issue. The combination of low sample sizes and convenience sampling limits external validity, meaning that the findings cannot be confidently generalized to broader populations of interest.

Second, the ascertainment of the exposure (Selection 4, M = 0.92, SD = 0.25; out of 2) is another area of concern. Studies routinely inadequately described or validated the measurement tools used, which can introduce measurement bias and affect the reliability of the results. Without validated tools, the accuracy and consistency of the data collected cannot be asserted with confidence, limiting internal validity.

Third, the comparability of groups (Comparability, M = 0.53, SD = 0.53; out of 2) is problematic across studies. This criterion assesses internal validity by evaluating whether the study controlled for confounding variables that could affect the results. Without proper model controls, and without experimental randomization, a study cannot confidently attribute observed effects to the intervention or variable of interest rather than to confounding factors.

Outcome assessment (Outcome 1, M = 1.97, SD = 0.14; out of 2) was excellent, but the appropriateness of statistical tests used (Outcome 2, M = 0.68, SD = 0.45; out of 1) had considerable issues. For outcome assessment, although many studies employed validated measurement tools, some relied on self-report measures, which can introduce bias and affect the reliability of the results. The appropriateness of statistical tests varied, with some studies using appropriate and clearly described statistical tests, including confidence intervals and *p*-values, while others used inappropriate or poorly described tests, making it challenging to draw valid conclusions from the data.

MMAT Summary Results. The MMAT evaluates the methodological quality by assessing qualitative, quantitative (randomized and non-randomized), and mixed methods studies, each with specific evaluation criteria. The results of the MMAT appraisal for the FSI articles are reported in Appendix Table C. Overall, the articles are of low quality, with a mean MMAT score of 3.20 (SD = 1.02), out of a possible 5 points.

For qualitative studies, the MMAT assessed the appropriateness of the qualitative approach, adequacy of data collection methods, derivation of findings from data, substantiation of results by data, and coherence between data sources, collection, analysis, and interpretation. These studies varied in quality—some met most evaluative criteria, while others failed to adequately justify their methodological choices.

Quantitative randomized controlled trials (IDs 9 and 18), self-identified as experiments, were evaluated based on proper randomization, baseline comparability of groups, completeness of outcome data, blinding of outcome assessors, and adherence to the assigned intervention. While these two studies were of an overall higher methodological quality compared to the rest of the FSI corpus, they nonetheless fell short of the standards required for reliable causal inference. Consequently, even the highest quality studies (by intended design) within the FSI corpus exhibit critical limitations that restrict their validity and applicability to evidence-based policy or judicial contexts.

Quantitative non-randomized studies (IDs 5, 8, 11, 13, 16, 20, and 24) were assessed on the representativeness of participants, appropriateness of measurements for outcomes and exposures, completeness of outcome data, control of confounders, and consistency in administering interventions. These studies lacked participant representativeness and controlling confounders, leading to lower overall scores.

Quantitative descriptive studies (IDs 2, 3, 4, 7, 10, 12, 14, 15, 17, 19, 21, and 22) were evaluated for the relevance of the sampling strategy, representativeness of the sample, appropriateness of measurements, low risk of nonresponse bias, and suitability of statistical analysis. These studies had varied performance, with some failing to ensure representative samples and appropriate measurement tools.

The sole mixed methods study (ID 23) was appraised based on the rationale for using mixed methods, effective integration of study components, adequate interpretation of integrated results, addressing inconsistencies between qualitative and quantitative findings, and adherence to quality criteria of both methods. This study had challenges in effectively integrating qualitative and quantitative components.

Specific Findings for Each Article Under Review

In addition to the overall critical appraisal, we assessed individual studies. Table 4 reviews ratings for each scale across the twenty-two scorable studies. The overall sum of scores, which could reach a maximum of 20, averaged 9.97 (SD = 3.61) across the studies. This suggests that *Force Science* research achieved less than half of the possible scientific reliability and quality benchmarks.

Notably, no single study scored highly across all three tools, underscoring a pervasive lack of methodological robustness, though study ID18 achieved the highest score relative to

Table 4. Summary Ratings Across Studies.

ID	MMAT sum (Max 5)	NOQ sum (Max 10)	MSMS (Design, Max 5)	MSMS (Implementation, Max 5)	Sum (20 possible)
2	1	3	2	1	7
3	2	3.33	0	0	5.33
4	2	3	0	0	5
5	3.33	4	1.33	0.67	9.33
7	4	6.67	2	1	13.67
8	3.67	5	2	1	11.67
9	3.33	4	5	4	16.33
10	5	5	2	1	13
11	4	5	2	1	12
12	3	5	0.67	0.33	9
13	3.67	5.33	2	1	12
14	2.67	5.33	0	0	8
15	2.67	4	0	0	6.67
16	3	4.67	2	1	10.67
17	2	4	0	0	6
18	3.33	4.67	5	4	17
19	2.67	2.33	0	0	5
20	5	4	2	1	12
21	2.67	5	2	1	10.67
22	3	2	0	0	5
23	3.33	5.67	2	1	12
24	5	4	2	1	12
Mean (SD)	**3.20 (1.02)**	**4.32 (1.12)**	**1.55 (1.44)**	**0.91 (1.10)**	**9.97 (3.61)**

other studies, scoring 17 out of a possible 20. This higher performance relative to the rest of the *Force Science* corpus was due to robust experimental design and reporting, though concerns about sample size (n = 76), lack of sample size justification, sample representativeness, and lack of statistical power detracted from the overall score.

This lack of methodological robustness raises serious questions about the scientific reliability of FSI's work. The observed variability and overall low scores imply that the research does not hold up to the standards required for credible scientific influence, particularly in legal contexts where formal adjudications hinge on validated, peer-reviewed evidence.

Findings for FSI-Defined Thematic Areas

This study also assessed the rigor of FSI findings grouped into the thematic areas as defined in *Force Science*. Figure 1 collects the numerical ratings across the thematic areas—

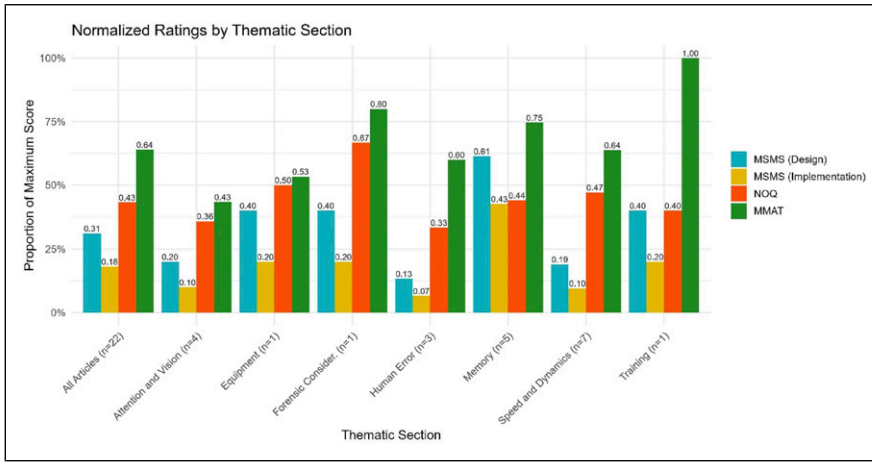

Figure 1. Normalized ratings by thematic section.

Foundational Principles, Attention and Vision, Memory, Speed and Dynamics, Forensic Considerations, Human Error, Equipment, and Training—and compares mean normalized scores (i.e., proportion of maximum possible) across these thematic sections for each metric (MMAT, NOQ, MSMS–Design, MSMS–Implementation). Overall, the figure illustrates a marked shortfall in meeting expected methodological standards. No section consistently achieved high scores across the metrics, with many studies rating only moderately or poorly on key dimensions of research design, sampling, and confounder control. See Appendix Table D for the full tabular results presented in Figure 1.

Figure 1 also indicates that certain thematic areas, such as Forensic Considerations and Training, exhibit relatively higher average scores. However, each of these sections encompasses only a single study, meaning these findings must be interpreted with caution. Single-study sections may appear stronger simply due to the absence of other, potentially weaker studies in that category. Consequently, these higher mean ratings do not necessarily reflect robust or consistently replicable research practices. Instead, they reflect the need for additional, high-quality research within these themes to determine whether the observed methodological adequacy is sustained across multiple studies.

More granularly, there was notable variation in methodological quality across thematic sections, with "Memory" emerging as relatively less flawed (compared internally to other FSI thematic areas) while "Attention and Vision" as well as "Speed and Dynamics" display considerable methodological deficiencies. These results imply that substantial improvements in research design, sampling strategy, and control of confounding factors would be necessary for FSI's studies to meet the methodological standards expected of scientific research influencing high-stakes legal and policy decisions.

Foundational Principles. The two studies in this section (ID1 and ID6) were excluded from scoring due to their status as narrative reviews rather than empirical research.

Consequently, no quantitative scores were assigned, as the assessment tools used in this review are designed for evaluating empirical studies. Unlike true scientific reviews, which synthesize evidence from peer-reviewed, methodologically rigorous studies, these pieces rely heavily on non-scientific sources, including magazine articles, blog posts, and other practitioner-oriented outlets. For example, study ID6 references at least fifteen non-scientific sources out of forty total.

Both reviews were published in *Law Enforcement Executive Forum*, a professional journal where Dr Lewinski is a member of an editorial board that includes non-researchers. The publication of these reviews in a journal where the author is a member of the editorial board raises concerns that they were not subject to impartial peer review or the critical scrutiny expected of academic science.

These concerns are reinforced when considering how often the publications engage in self-citation to Dr Lewinski's own work, and how those works were not themselves placed in scientific outlets. For example, study ID1 relies on citations to four blog posts or articles published in the non-scientific magazine *Police Marksman*, of which list Dr Lewinski as each listing Dr Lewinksi as the first author. A similar pattern emerges in study ID6, which relies on citations to a total of eight blog posts, magazine articles, and a publication in *LEEF* (as noted throughout, a non-scientific practitioner outlet), all of which identify Dr Lewinski as the first author.

These two articles lack the hallmarks of scientific reliability. Scientific review articles rarely, if ever, lean so heavily on non-scientific sources or unpublished materials. As such, these 'reviews' should be interpreted with caution, as they do not meet the evidentiary standards of independent, empirically grounded research.

Attention and Vision. The four articles on "Attention and Vision" (IDs 2, 3, 4, and 8) exhibited serious methodological gaps, with most studies failing to meet higher standards of research rigor. The MMAT averaged a low score of 2.17 (SD = 1.11), suggesting substantial limitations in research design. The NOQ was similarly low (M = 3.58, SD = 0.96), underscoring issues with sample selection and the control of confounding factors. MSMS design scores were minimal (M = 1.0, SD = 1.15), indicating weak causal inference capacity, and issues with implementation drve scores down (M = 0.5, SD = 0.58).

Memory. The five articles addressing "Memory" (IDs 5, 9, 13, 18, and 20) demonstrated relatively stronger methodological quality, as reflected in an average MMAT score of 3.73 (SD = 0.72) and a NOQ score of 4.4 (SD = 0.6). These scores suggest that studies in this section achieved moderate quality, bolstered by a relatively higher methodological focus on data consistency. MSMS scores for design (M = 3.07, SD = 1.79) and implementation (M = 2.13, SD = 1.71) were the highest among all sections, but overall still fell short of the expected level of methodological rigor for evidentiary reliability.

Speed and Dynamics. The eight articles in the "Speed and Dynamics" section (IDs 10, 11, 12, 14, 15, 16, and 17) scored variably across measures, with an MMAT average of 3.19

(SD = 1.0) and a NOQ average of 4.71 (SD = 0.52), indicating moderate quality overall but with variability in rigor. Notably, MSMS design (M = 0.95, SD = 1.01) and implementation (M = 0.48, SD = 0.5) scores were particularly low, suggesting pervasive issues with study design, implementation, and the validity of causal inferences. These findings raise critical questions about the robustness of scientific claims made in this domain, particularly in studies that intend to inform tactical and performance-based recommendations.

Human Error. The three articles in the "Human Error" section (IDs 19, 22, and 23) demonstrate substantial limitations, as studies in this domain received low scores for both design integrity and implementation quality. The studies revealed consistently low methodological rigor, with an MMAT score of 3 (SD = 0.33) and an average NOQ score of 3.33 (SD = 2.03). MSMS scores were similarly low, with averages of 0.67 (SD = 1.15) for design and 0.33 (SD = 0.58) for implementation, reflecting serious methodological gaps.

Forensic Considerations, Equipment, and Training. Three sections contain only a single article: Forensic Considerations, Equipment, and Training. While they contribute to pooled analyses in the summary results, the solo nature of these sections prevents deeper thematic analysis. The single article in the "Forensic Considerations" section (ID (7) scored higher in methodological quality than most sections. With a high MMAT score of 4 and a high NOQ score of 6.67, this study exhibited strong methodological rigor. MSMS scores for design (M = 2) and implementation (M = 1) reflect the lack the ability to demonstrate causal relationships, a common issue across the *Force Science* corpus. Overall, the scores suggest a stronger empirical basis relative to the FSI corpus, though conclusions about overall reliability remain tentative given there is only a single article.

The single article in the "Equipment" section (ID 21) attained a mean MMAT score of 2.67, with a NOQ score of 5 and modest MSMS ratings (M = 2 for design, M = 1 for implementation). We cannot infer across a single article, but our analysis does not indicate high scientific reliability for this article.

The single study in the "Training" section (ID 24) scored a perfect 5 on the MMAT, indicating a solid methodological foundation, although the NOQ score was an average 4, suggesting some limitations when compared to other designs. The same pattern of low MSMS scores across design (2) and implementation (1) appear. All three sections have only a single article each; the limited number of studies tempers the generalizability of these findings.

Discussion

To our knowledge, this is the first study to systematically review the articles that FSI collected to present its most relevant theories and concepts. *Force Science* presents the collected articles and their related concepts as scientifically rigorous, thereby justifying their admissibility in legal proceedings and use in formulating police training, policies, and behavior. However, when judged against multiple benchmarks for scientific research, the articles collected in *Force Science* have considerable gaps and

weaknesses. These findings raise serious concerns about the lack of scientific reliability and validity of FSI findings, suggesting the need for caution when considering the use of these materials in policy or legal contexts These concepts simply do not derive from an appropriately rigorous body of scientific research.

Our review does not assess whether the conclusions presented in Force Science publications are empirically false, but instead finds insufficient scientific grounding to assert they are empirically true. Establishing such claims would require conclusions derived from rigorous, methodologically sound research. At the individual level, and especially when used to inform legal decision-making or judicial verdicts, FSI's assertions extend well beyond what the available data can support.

Scientific Rigor and Admissible Evidence

Under *Daubert*, evidentiary reliability demands internal and external validity to establish scientific credibility. The lead author of study ID18 (the highest-scoring article among the *Force Science* studies) articulates in a separate review of stress and fatigue effects on police response and memory recall (Hope, 2016, p. 242) that"

> …applied research must be held to the same high methodological standards as laboratory work, including (i) recruitment of adequate samples sizes to test well-defined, theory-driven hypotheses; (ii) inclusion of relevant control groups; (iii) implementation of controlled, replicable scenarios; (iv) systematic manipulation of independent variables; and (v) use of appropriate statistical analyses. We endorse these principles. For empirical research to qualify as scientifically credible, validity is established through high ratings across methodological quality measures that assess internal and external validity, such as the MMAT, NOQ, and MSMS.

When we apply those principles and assessment tools, we find that, as a whole, *Force Science* fails to meet fundamental scientific standards. This pattern highlights persistent methodological weaknesses that permeate the individual *Force Science* studies. Although a few studies demonstrated specific strengths, as reflected in moderate MMAT scores or isolated NOQ components, these same studies lacked rigor in other fundamental areas, including sampling strategies, validation of outcome measures, and confounder control. This pattern indicates the FSI's need to significantly strengthen the rigor of its research designs, especially regarding sampling and procedural controls, to satisfy the scientific standards required for credible findings.

That only a few studies scored even moderately well indicates that the current body of work does not meet a quality threshold that establishes it as internally or externally valid. Therefore, the studies lack the reliability necessary for admissibility in legal proceedings and, by extension, their suitability in policy contexts.

The *Daubert* criteria—testability, error rate, general acceptance, and peer review and publication—provide the framework for evaluating the admissibility of scientific expert testimony. Fournier (2016) outlines, from a research perspective, how those criteria

ensure a minimum quality of scientific research presented in court. Fournier (2016, p. 309) asserts that excluding expert opinions that lack an adequate scientific basis is essential to reducing unreliable, biased testimony that may inappropriately influence a finding or verdict. Here, we take up the *Force Science* articles in terms of these criteria:

Testability

Testability is a cornerstone of scientific integrity and a key aspect of the *Daubert* standard. Many FSI studies exhibit limited testability due to reliance on small convenience samples, a flaw also reflected in their low MSMS design scores. Without random sampling or adequately powered sample sizes, generalizability to diverse policing contexts cannot be established.

Study ID12, for example, claims its results "can be directly correlated to an average officer's simple perception and movement time" in a trigger-pull measurement (Lewinski et al., 2014, p. 11). However, this study relied on a small convenience sample (n = 108) with no effort to report complete sample characteristics (only a gender breakdown is given). This lack of adequate sample reporting is repeated in study ID2 (Lewinski, 2008), ID3 (Schwarzkopf et al., 2008), ID5 (Page et al., 2008), ID15 (Lewinski et al., 2015), and ID23 (O'Neill et al., 2021).

This lack of generalizability, compounded by a lack of power analyses or sample size justification, undermines the ability to draw reliable conclusions and apply findings in new contexts, such as when an expert attempts to apply scientific principles in a new case under legal scrutiny. These limitations weaken the credibility and reliability of FSI's scientific contributions, especially when their findings are intended to inform critical judicial decisions and law enforcement training.

Error Rates

Error rates constitute a considerable shortcoming in FSI studies, with reviewed articles at times omitting key metrics for assessing the precision of findings. For example, in study ID12, which purports to report on average police officer reaction time to fire a training handgun, the narrative results report mean times without consideration of an error term and no sample statistics are reported.

There is also often no effort to control for subject characteristics. In study ID8 (Vickers & Lewinski, 2012), which explores what officers look at, the authors acknowledge a significant difference in age between their "elite" and "rookie" officers, although no effort is made to adjust the results statistically for this difference, or to address potential confounders. Similarly, in ID12 (Lewinski et al., 2014), no effort is made to model the potential confounders of gender, age, or experience. This is a common issue in our review of *Force Science*, and this gap is further compounded by the frequent use of small, non-random convenience samples, which introduces selection bias and restricts the generalizability of findings.

These limitations are evident in the studies' low NOQ and MMAT scores, reflecting a lack of statistical rigor and transparency. Without appropriate reporting and model controls, readers—particularly in non-experts in legal contexts—cannot adequately assess the influence of random error or selection bias on the results, nor can they determine whether the findings are applicable beyond the immediate study sample.

General Acceptance

General acceptance within the scientific community is another component of the *Daubert* standard in which FSI studies suffer from substantial limitations. From a high level, an important measure of "general acceptability" is how the wider scientific community would evaluate the scientific rigor of the studies underlying the expert's claim of scientific evidence. As we've shown, the *Force* Science corpus does not comport with well-accepted scientific standards.

As a corollary inquiry, well-accepted scientific research typically demonstrates its value through significant engagement by other scholars, including citations and integration into subsequent research. The FSI studies reviewed here show very limited scholarly engagement. While this lack of engagement could partially reflect limited interest in their chosen topics, it more likely stems from the methodological weaknesses and outlet preferences outlined throughout this paper. Consistently low scores reveal persistent departures from basic scientific design and execution, undermining the credibility of *Force Science* and making it less likely that other experts in the field would rely on these works to the extent required by the *Daubert* framework.

Peer Review & Publication Standards

Peer review and publication standards are central to enforcing scientific norms of internal validity and reliability. Chan (1995) draws a critical distinction between "true peer review" and "editorial peer review," terms frequently conflated but with significantly different implications for scientific validity. True peer review involves rigorous, ongoing scrutiny, and replication by the broader scientific community, representing a continuous process extending beyond initial publication. In contrast, editorial peer review refers to the preliminary, discretionary assessment conducted by journals to decide whether a manuscript merits publication, focusing primarily on methodological soundness, originality, and relevance, rather than comprehensive validation of scientific claims. as they do not represent, or claim to represent, genuine *scientific* evaluation. Scientific peer review must involve reviewers who possess the necessary methodological expertise to assess empirical validity. LEEF's publication practices fall short even of Chan's more permissive category of editorial peer review, further undermining any claim to scientific legitimacy.

As Fournier (2016) points out, these scientific qualities are necessary for the relevance of studies in legal and policy frameworks. The *Daubert* court (p. 594) emphasized the importance of scientific peer review, as (emphasis added) "submission to

the scrutiny of the *scientific community* is a component of 'good science,' in part because it increases the likelihood that substantive flaws in methodology will be detected." *Daubert* recognized that scientific peer review is not a cure-all. Even high-ranking scientific journals have struggled with replication, with studies suggesting that only one-third to one-half of findings are consistently reproducible (Open Science Collaboration, 2015). Recognizing this, many disciplines have embraced open science practices such as study pre-registration to address threats to scientific credibility (Nosek et al., 2018). Nevertheless, high-quality peer review remains a cornerstone of ensuring methodological rigor and scientific integrity, especially when research is intended to inform legal and policy frameworks.

This component of the *Daubert* framework exposes perhaps the starkest deficiency in the scientific credibility of the *Force Science* studies. The bulk (41.6%) of the *Force Science* articles reviewed here were published in the *Law Enforcement Executive Forum* (*LEEF*), a non-scientific, practitioner-focused publication. *LEEF* makes no attempt to hold itself out as a scientific publication. The exclusion of *LEEF* from Web of Science further signals its status as a professional, rather than a scientific, outlet. This classification reflects its lack of qualities indicative of rigorous empirical research: comprehensive scientific peer review, robust publication practices, well-defined ethical standards, and a citation record that reflects its impact on the field (Clarivate, 2024). Even accepting Chan's more permissive category of editorial peer review as a minimal baseline, LEEF, where more than 40% of the Force Science corpus appears, fail to meet that threshold as they do not represent, or claim to represent, genuine *scientific* evaluation. LEEF's publication practices fall short even of Chan's more permissive category of editorial peer review, further undermining any claim to scientific legitimacy. This context emphasizes why academic scientists very rarely cite *LEEF* as a source of scientifically credible evidence.

Scholars routinely translate their peer-reviewed findings into concise, practitioner-facing pieces (e.g., in *Police Chief* or the *FBI Law Enforcement Bulletin*) so that field personnel can apply the vetted evidence without wading through scientific journals. Because these summaries explicitly cite, and thus preserve the evidentiary chain to, the original refereed studies, they complement rather than replace formal scientific review. The heavy reliance on *LEEF* to disseminate FSI's findings in the midst of a research community that publishes a wide range of more rigorous journals suggests that publication in *LEEF* could be characterized as a strategic decision not to interpret scientifically peer-reviewed infromation for practitioners, but to bypass the scrutiny of more scientifically demanding journals altogether.

Toward a Science of Force

There is a pressing need for reliable knowledge about the physiological and behavioral dynamics that can apply in use-of-force situations. Such information is essential not only for helping courts reach fair and consistent verdicts, especially amid heightened scrutiny of police practices, but also for guiding upstream decisions in legal processes

and informing police department policies, procedures, and training that promote the proper use of force in the first place. If nothing else, broad acceptance of FSI as a source of expertise in policing and the legal arena indicates a great thirst for this knowledge in the midst of what is, for the most part, a vacuum.

It should be stressed that the articles in the *Force Science* volume studied here are scientifically valuable precisely because of their methodological weaknesses. They provide a roadmap toward a science of force, even if they cannot begin to settle the empirical questions they attempt to answer. Closely examining those weaknesses will help identify research questions of interest, provide data for formulating hypotheses, and suggest possible study designs, all of which need to be considerably expanded and improved before any findings could be admissible under *Daubert.*

In making such progress, researchers conducting studies in this area should prioritize independence and transparency to ensure the credibility of their findings. Organizations commissioning or benefiting from this research should support independent investigators, avoid direct involvement in study design or execution whenever possible, clearly declare the obvious conflicts of interest when they do not, and rely on impartial results to guide their practices. To enhance the quality and impact of the research, authors should subject their work to scientific peer review, avoid publishing in obscure journals with nonstandard review processes, and commit to open science practices including pre-registration and open data and replication code (Chin et al., 2023). The inherent tension between organizations with financial or operational stakes in contested, emerging research topics and the scientific imperative for impartial, rigorous study underscores the need for stronger safeguards to ensure objectivity and credibility in this field.

Limitations

Our analysis has limitations. First, our sample is neither a random nor complete sample of FSI's research articles. We did not draw our sample from the nebulous and undefined entirety of the FSI-influenced corpus, but instead focused on the two dozen articles selected by FSI itself for collection in *Force Science*. While Lewinski (2022, p. XV) claims that "well over a thousand peer-reviewed and professional journal articles" have been published by affiliates of FSI, there is no comprehensive accounting of this body of claimed ancillary research, making it impossible to either validate his claim or evaluate the rigor of that ostensibly sprawling corpus of work. It seems likely that, whatever their number, the vast majority of these ancillary publications are professional articles and gray literature unencumbered by double-blind peer review processes or requirements of scientific rigor. Regardless, FSI collected the twenty-four articles in *Force Science* that, in its own estimation, represent "some of the most critical questions asked by the Force Science research team" (p. XV). FSI holds these twenty-four articles up as "peer reviewed scientific research," emphasizing that description by way of the book's subtitle, and the preface underlines the explicit goal of influencing both law enforcement and the courts. Thus, while we did not analyze the entire corpus of FSI-

related publications to select the research we reviewed, our evaluation focuses on the articles that FSI itself characterizes as the most representative and rigorous of its scientific research.

While the methodological assessment tools employed in this study—the Maryland Scientific Method Scale (MSMS), Newcastle-Ottawa Quality Assessment Scale (NOQAS), and the Mixed Methods Appraisal Tool (MMAT)—are well-established and accepted, each carries inherent limitations that warrant caution in interpretation. First, the Maryland Scale primarily emphasizes causal inference and research design hierarchy, thereby systematically favoring studies employing experimental or quasi-experimental designs. This emphasis limits its sensitivity to adequately distinguish the methodological rigor of descriptive or exploratory research. Consequently, it may undervalue certain studies that appropriately employ non-experimental methods to address research questions unsuited to causal inference frameworks. Second, the Newcastle-Ottawa Quality Assessment Scale, adapted here for cross-sectional studies, prioritizes clear reporting and representativeness, yet it places significant weight on sample size justification and representativeness of the sample. As such, it might disproportionately penalize studies conducted under resource or practical constraints typical in specialized or applied settings, such as policing research. Finally, the Mixed Methods Appraisal Tool, while broadly applicable across diverse research designs, offers only a binary scoring system (criteria met vs. not met), potentially over-simplifying nuanced methodological strengths and weaknesses within complex studies. However, in using each of these three different tools, our approach helps overcome the weaknesses inherent to any one scoring mechanism.

Another limitation is that we may have mischaracterized *Force Science* itself. We approach our critical appraisal with the perspective that the volume as purports to consist of two dozen rigorous articles that convey high quality evidence in an established scientific field. One response might be that the articles instead represent pioneering, formative work in a nascent, immature field that is setting the stage for a future arc of experimentation. If that is the case, the articles would be characterized as pilot studies and preliminary data that provide valuable bases for hypothesis formation, increasingly rigorous study designs, and a means to adapt and refine experimental methods. This would provide an explanation for the lack of reliance by other scholars, the comparatively low citation counts, and low (or no) impact journals that house *Force Science* publications. It would also explain why the study designs and sampling were not typical of those seen in large, high-quality studies, but rather ones that attempt to work out the concepts and approaches that justify such studies. In other words, *Force Science* may provide us with the useful seeds of a science of force without attempting to do more. While this is not how FSI portrays its own research, perhaps their stridency in asserting it is well-established science is an understandable feature of researchers' natural faith in and enthusiasm for their own work. Regardless, this point would decisively settle the principal concern of our paper: such nascent, pioneering work could not be admitted as reliable evidence in a legal proceeding or the formulation of

policies and procedures. In other words, even as pioneering research, *Force Science* decisively fails the *Daubert* test.

Conclusion

As of April 2025, under the "Research" section the FSI website endorses the value of what Chan (1995) calls "true peer review," stating: "We recognize that science is an evolving pursuit that demands continual critical examination. Research findings must be tested, challenged, and refined as new data and methodologies emerge" (Force Science Institute, 2025). Our review undertakes precisely that, critically evaluating the Force Science corpus in terms of adherence to established scientific norms and methodological rigor.

Our review reveals that *Force Science* studies have significant methodological shortcomings. The volume's studies consistently fail to meet prevailing scientific norms, rendering it unreliable as an evidentiary matter under *Daubert*. The underlying FSI concepts have nevertheless been forced into the courtroom, providing a veneer of scientific credibility to analysis that rests on an inadequate foundation. This lack of rigor undermines the credibility of the findings in *Force Science*, indicating they are unsuitable for use in legal settings or as a foundation for shaping police policy and practice.

The implications of this review extend beyond a detached academic critique. Police training, policies, and judicial decisions justified by methodologically weak research risk perpetuating unnecessarily harmful practices that could otherwise be improved or remediated through reliance on more rigorous studies, a possibility that judicial acceptance of weaker, less substantiated research findings has the practical effect of foreclosing (Fournier, 2016; Reisberg et al., 2016). This raises deep moral concerns about the use of potentially unreliable, if putatively scientific, claims in contexts where life and liberty hang in the balance, eroding the legitimacy of institutions that rely on this evidence to underwrite their highly consequential judgments. FSI's research addresses crucial questions in law enforcement. However, the pervasive lack of rigor identified in this analysis shows their findings cannot reliably inform policy or courtroom outcomes. We emphasize the urgent need for methodologically robust research in police practices and officer performance, conducted with the rigor necessary to establish scientific advice that can achieve evidentiary reliability in the legal arena.

These findings also have implications for evidence-based policing, an approach emphasizing policies and practices grounded in rigorous scientific research. Scholars have argued that reliable evidence is essential for developing policing policies that promote both effectiveness and accountability (Lum & Koper, 2017). Further, police training in high-risk areas that are routinely scrutinized in court should likewise be grounded in scientifically robust evidence to prevent the entrenchment of dubious methods or practices. When research lacks rigor, it not only jeopardizes the quality of decisions that may directly affect the public, it also leaves open the risk of police training and policies formulated under false or unnecessary assumptions. If that is the case, then it is possible that an overreliance on the precepts found in *Force Science* has

precipitated use of force incidents that the FSI then attempts to justify, while appearing to foreclose the scientific possibility of evidence-based training and policies that would reduce the necessity of these uses of force in the first place. Consequently, the stakes of endorsing insufficiently substantiated claims are high, both for the individuals whose rights are at risk and for the credibility of law enforcement agencies tasked with upholding those rights. The methodological weaknesses observed across *Force Science* show that it lacks the scientific rigor required to reliably inform court proceedings, evidence-based policing, or police policy and training.

Acknowledgement

A preliminary version of this article was presented at the American Society of Evidence-Based Policing (ASEBP) conference, May 2025, at the University of Arizona. We thank ASEBP President Jason Potts for rejecting an effort by the Force Science Institute to interfere with our presentation. In his response to Dr. Bill Lewinski and the Force Science Institute (April 7, 2025), Chief Potts wrote: "We do not reject or alter presentation content based on its viewpoint, nor do we censor titles or topics, even if they are critical of specific research bodies, methodologies, or institutions. ASEBP believes such oversight would be antithetical to the scientific process and to the values of transparency, critical peer review, and progress through debate…ASEBP does not, and will not, interfere with the intellectual autonomy of our presenters. We view their contributions as part of a collective endeavor to improve public safety outcomes, elevate the quality of research in policing, and respect the constitutional and academic rights of our participants." We fully endorse this stance and appreciate ASEBP's commitment to academic freedom. In the spirit of academic inquiry, we also appreciate and concur with the decision of Dr. John Worrall, Editor-in-Chief of Police Quarterly, to invite Force Science Institute to respond to our paper, which they were provided in advance of publication.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Dr. Adams, Dr. Alpert, and Professor Stoughton disclose receiving fees and/or reimbursements as expert witnesses in civil and criminal litigation. In that context, they have worked with and been adversarial to experts affiliated with, or relying on concepts promulgated by, the Force Science Institute. Dr. Adams, Dr. Alpert, Professor Stoughton, and Irick Geary disclose receiving fees and/or reimbursements for providing police training on topics that may include those of interest in this study, including use of force. Irick Geary discloses that, in his professional capacity, he makes decisions about use-of-force training, vendors, and experts that overlap with concepts related to those promulgated by the Force Science Institute. Dr. Adams and Irick Geary disclose having participated in training offered by the Force Science Institute.

Funding

Dr. del Pozo was supported by the National Institute on Drug Abuse (grant K01DA056654). The Institute had no role in the conduct of this research or the preparation of this manuscript, which

may not reflect its policies or positions. The remaining authors report no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Ian T. Adams https://orcid.org/0000-0001-5595-8070
Seth Stoughton https://orcid.org/0000-0002-1873-1333
Brandon del Pozo https://orcid.org/0000-0001-6481-2196
Irick T. J. Geary https://orcid.org/0009-0002-5144-8959
Marc Olson https://orcid.org/0009-0006-8479-7003

Supplemental Material

Supplemental material for this article is available online.

Notes

1. 293F. 1013 (D.C. Cir. 1923).
2. 509 U.S. 579 (1993).
3. We recognize that the *Daubert* standard applies in federal courts, while some state jurisdictions apply the *Frye* standard or modified versions of *Daubert*. However, given that many high-profile police use-of-force cases are litigated federally under Section 1983 claims, Daubert remains particularly relevant. Thus, our focus remains explicitly federal. Future research may benefit from examining state-level variations in admissibility standards and their implications for policing expertise and evidentiary practices.
4. Our analysis specifically evaluates the scientific rigor and methodological quality of studies published by the Force Science Institute (FSI). Courts may also admit testimony based on non-scientific expertise—such as technical or specialized knowledge—under standards articulated by the Supreme Court in *Kumho Tire Co. v. Carmichael* (1999). However, our critique remains strictly focused on the scientific validity and reliability of Force Science claims, as these claims are frequently presented and defended explicitly as scientific evidence.
5. As noted before, two articles (IDs 1 and 6) are non-empirical review pieces, and were not reviewed using the appraisal tools.
6. *In re Wholesale Grocery Prods. Antitrust Litig.*, 946 F.3d 995, 1001 (8th Cir. 2019).

References

Apuzzo, M. (2015). Training officers to shoot first, and he will answer questions later. *The New York Times*. https://www.nytimes.com/2015/08/02/us/training-officers-to-shoot-first-and-he-will-answer-questions-later.html

Blake, D. (2013). How to apply Force Science findings to policy and training. *Police1*. https://www.police1.com/officer-safety/articles/how-to-apply-force-science-findings-to-policy-and-training-ccBz4tJkQKvbPquB/

Brown, A. (2024). CT trooper on trial acted "reasonably," experts say – but not in court. *CT Mirror*. https://ctmirror.org/2024/03/12/ct-trial-brian-north-mubarak-soulemane-shooting/

Chan, E. J. (1995). The Brave New World of Daubert: True peer review, editorial peer review, and scientific validity. *New York University Law Review*, *70*(1), 100. https://heinonline.org/HOL/Page?handle=hein.journals/nylr70&id=116&div=&collection=

Chin, J. M., Pickett, J. T., Vazire, S., & Holcombe, A. O. (2023). Questionable research practices and open science in quantitative criminology. *Journal of Quantitative Criminology*, *39*(1), 21–51. https://doi.org/10.1007/s10940-021-09525-6

Clarivate. (2024). *Journal evaluation process and selection criteria*. https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/web-of-science-core-collection/editorial-selection-process/journal-evaluation-process-selection-criteria/

DeJong, C., & St George, S. (2018). Measuring journal prestige in criminal justice and criminology. *Journal of Criminal Justice Education*, *29*(2), 290–309. https://doi.org/10.1080/10511253.2017.1398344

Faigman, D. L. (2012). The Daubert revolution and the birth of modernity: Managing scientific evidence in the age of science. *UC Davis Law Review*, *46*(3), 893. https://heinonline.org/HOL/Page?handle=hein.journals/davlr46&id=913&div=&collection=

Faigman, D. L., Kaye, D. H., Saks, M. J., & Sanders, J. (1999). How good is good enough: Expert evidence under Daubert and Kumho. *Case Western Reserve Law Review*, *50*(1), 645. https://heinonline.org/HOL/Page?handle=hein.journals/cwrlrv50&id=665&div=&collection=

Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2002). The Maryland scientific methods scale. In *Evidence-based crime prevention* (pp. 13–21). Routledge.

Force Science. (2018a). *About us*. https://www.forcescience.com/about/

Force Science. (2018b). *Consulting—force science*. https://www.forcescience.com/consulting/

Force Science. (2019). *Upcoming training events listing*. https://www.forcescience.com/training/events/?Event=Force_Science_Certification

Force Science. (2022a). *Force encounters course | investigations & human performance*. https://www.forcescience.com/training/force-encounters-course/

Force Science. (2022b). *Force science certification course | evidence-based training*. https://www.forcescience.com/training/force-science-certification-course/

Force Science. (2023). *Advanced force science specialist course—100% online*. https://www.forcescience.com/training/advanced-force-science-specialist-course/

Force Science. (2024). *Intake form*. https://forms.forcescience.com/forcescience/form/ConsultingIntakeForm/formperma/izFZ6aDwiz5TJPrDWyc5wo8xqDs-62HD8nmej-WbPs8

Force Science Institute. (2025). Peer-reviewed research—force science. *Commitment to Scientific Integrity and Open Inquiry*. https://www.forcescience.com/research/

Fournier, L. (2019). *Expert report on Jamie Borden* (Declaration Nos. 2: cv-18–00055; Issues 2: cv-18–00055). U.S. District Court for the Western District of Washington.

Fournier, L. R. (2011). *Fournier peer review of Lewinski articles and dissertation* (Declaration of Dr. Fournier Nos. 9-CR-0088-FVS). United States District Court.

Fournier, L. R. (2016). The Daubert guidelines: Usefulness, utilization, and suggestions for improving quality control. *Journal of Applied Research in Memory and Cognition*, *5*(3), 308–313. https://doi.org/10.1016/j.jarmac.2016.06.012

Gatowski, S. I., Dobbin, S. A., Richardson, J. T., Ginsburg, G. P., Merlino, M. L., & Dahir, V. (2001). Asking the gatekeepers: A national survey of judges on judging expert evidence in a post-Daubert world. *Law and Human Behavior*, *25*(5), 433–458. https://doi.org/10.1023/A:1012899030937

Giannelli, P. C. (1980). The admissibility of novel scientific evidence: Frye v. United States, a half-century later. *Columbia Law Review*, *80*(6), 1197. https://doi.org/10.2307/1122061. https://heinonline.org/HOL/Page?handle=hein.journals/clr80&id=1215&div=&collection=

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

Hong, Q. N., Gonzalez-Reyes, A., & Pluye, P. (2018a). Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). *Journal of Evaluation in Clinical Practice*, *24*(3), 459–467. https://doi.org/10.1111/jep.12884

Hong, Q. N., & Pluye, P. (2019). A conceptual framework for critical appraisal in systematic mixed studies reviews. *Journal of Mixed Methods Research*, *13*(4), 446–460. https://doi.org/10.1177/1558689818770058

Hong, Q. N., Pluye, P., Fabregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018b). *MMAT criteria manual*.

Hope, L. (2016). Evaluating the effects of stress and fatigue on police officer response and recall: A challenge for research, training, practice and policy. *Journal of Applied Research in Memory and Cognition*, *5*(3), 239–245. https://doi.org/10.1016/j.jarmac.2016.07.008

Huber, P. W. (1993). *Galileo's revenge: Junk science in ihe courtroom*. Basic Books.

Hyman, I. (2022). When expert witnesses are the problem. *Psychology Today*. https://www.psychologytoday.com/us/blog/mental-mishaps/202201/when-expert-witnesses-are-the-problem

Illinois Law Enforcement Training and Standards Board Executive Institute. (2015). Volume 15, Issue 1 of law enforcement executive forum. Law Enforcement Executive Forum, 15.

Illinois Law Enforcement Training and Standards Board Executive Institute. (2024). *Illinois law enforcement training and standards board executive institute*. ILETSBEI. https://iletsbei.org/

Kliem, V. (2020). Leading the national discussion on policing—force science. *Force Science News*. https://www.forcescience.com/2020/06/leading-the-national-discussion-on-policing/

Kliem, V. (2022). Experts work with force science to advance police-related research. *Force Science News*. https://www.forcescience.com/2022/08/top-experts-work-with-force-science-to-advance-police-related-research/

Kliem, V. (2023). Trainers as police practice and human factors experts—force science. *Force Science News*. https://www.forcescience.com/2023/01/trainers-as-police-practice-and-human-factors-experts/

Kliem, V. (2024). Aligning research on human performance across high-stakes professions—force science. *Force Science News*. https://www.forcescience.com/2024/09/aligning-research-on-human-performance-across-high-stakes-professions/

Krampl, A. (2019). Journal citation reports. *Journal of the Medical Library Association*, *107*(2), 280–283. https://doi.org/10.5195/jmla.2019.646

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lewinski, W. (2008). The attention study: A study on the presence of selective attention in firearms officers. *Law Enforcement Executive Forum*, *8*(6), 107–139.

Lewinski, W. (2010). *Testimony in the case of State of California v. Johannes Mehserle*. Los Angeles County District Court.

Lewinski, W. (Ed.), (2022). *Force science: Peer-reviewed scientific research*. XanEdu.

Lewinski, W., Dysterheft, J., Bushey, J., & Dicks, N. (2015). Ambushes leading cause of officer fatalities - when every second counts: Analysis of officer movement from trained ready tactical positions. *Law Enforcement Executive Forum*, *15*(1), 1–15.

Lewinski, W., Hudson, B., & Dysterheft, J. L. (2014). Police officer reaction time to start and stop shooting: The influence of decision-making and pattern recognition. *Law Enforcement Executive Forum*, *14*(2), 1–16.

Lum, C., & Koper, C. S. (2017). *Evidence-based policing*. Oxford Univ. Press.

Lvovsky, A. (2017). The judicial presumption of police expertise. *Harvard Law Review*, *130*(8), 1997. https://harvardlawreview.org/print/vol-130/the-judicial-presumption-of-police-expertise/

Madaleno, M., & Waights, S. (2016). *Quick scoring guide for the Maryland scientific methods scale*. What Works Centre for Local Economic Growth. https://whatworksgrowth.org/wp-content/uploads/Quick_Scoring_Guide.pdf

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/j.joi.2018.09.002

Modesti, P. A., Reboldi, G., Cappuccio, F. P., Agyemang, C., Remuzzi, G., Rapi, S., Perruolo, E., Parati, G., & ESH Working Group on CV Risk in Low Resource Settings. (2016). Newcastle-Ottawa quality assessment scale (adapted for cross sectional studies). *PLoS One*, *11*(1), Article e0147601. https://doi.org/10.1371/journal.pone.0147601

Muller v. Oregon. (1908) 208 US 412 (Supreme Court 1908).

Nave, C., Meehan, A. J., & Dennis, A. M. (2024). "You don't need a rocket scientist to figure out what could happen": Reasoning practices in police use of force trials. *Law & Social Inquiry*, *49*(4), 2439–2465. https://doi.org/10.1017/lsi.2024.19

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

O'Neill, D. A., Spence, W. R., Lewinski, W., & Novak, E. J. (2021). Training and safety: Potentially lethal blue-on-blue encounters. *Police Practice and Research*, *22*(2), 1209–1228. https://doi.org/10.1080/15614263.2019.1617143

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Page, J. W., Thibeault, C. M., Page, K. F., & Lewinski, W. (2008). Pursuit driver training improves memory for skill-based information. *Police Quarterly*, *11*(3), 353–365. https://doi.org/10.1177/1098611107310315

Plymouth County District Attorney's Office. (2021). *Findings of Plymouth county district attorney Timothy J. Cruz regarding officer-involved fatal shooting on February 16, 2020 in Byfield, Massachusetts*. Plymouth County District Attorney's Office.

Ratcliffe, J. H. (2022). *Evidence-based policing: The basics* (1st ed.). Routledge.

Rector, K. (2023). California attorney general clears LAPD officer in shooting using controversial "expert". *Los Angeles Times*. https://www.latimes.com/california/story/2023-01-12/ag-says-lapds-toni-mcbride-was-justified-wont-be-charged-in-daniel-hernandezs-killing

Reisberg, D., Simons, D. J., & Fournier, L. R. (2016). Introduction to the forum on when and whether psychological research is ready for use in the justice system. *Journal of Applied Research in Memory and Cognition*, *5*(3), 233–235. https://doi.org/10.1016/j.jarmac.2016.07.009

Remsberg, B. (2011). *Case studies: How force science analysts helped accused officers - force science*. Force Science News. https://www.forcescience.com/2011/09/case-studies-how-force-science-analysts-helped-accused-officers/

Schwarzkopf, E., Houlihan, D., Kolb, K., Lewinski, W., Buchanan, J., & Christenson, A. (2008). Command types used in police encounters. *Law Enforcement Executive Forum*, *8*(2), 99–114. https://cornerstone.lib.mnsu.edu/psyc_fac_pubs/21

Smith, M. R., & Alpert, G. P. (2002). Searching for direction: Courts, social science, and the adjudication of racial profiling claims. *Justice Quarterly*, *19*(4), 673–703. https://doi.org/10.1080/07418820200095391

Valentino-DeVries, J., McIntire, M., Ruiz, R. R., Tate, J., & Keller, M. H. (2021). How paid experts help exonerate police after deaths in custody. *The New York Times*. https://www.nytimes.com/2021/12/26/us/police-deaths-in-custody-blame.html

Vickers, J. N., & Lewinski, W. (2012). Performing under pressure: Gaze control, decision making and shooting performance of elite and rookie police officers. *Human Movement Science*, *31*(1), 101–117. https://doi.org/10.1016/j.humov.2011.04.004

Walrdon, J. (2011). The rule of law and the importance of procedure. In *Getting to the Rule of law* (pp. 3–31). New York University Press. https://doi.org/10.18574/nyu/9780814728437.003.0001

Author Biographies

Ian T. Adams, PhD, is an Assistant Professor of Criminology & Criminal Justice, and the Senior Research Advisor of the Excellence in Policing & Public Safety (EPPS) Program at the Joseph F. Rice School of Law, both at the University of South Carolina. He holds both basic and advanced certifications from the Force Science Institute.

Seth Stoughton is a Professor of Law & Faculty Director of the Excellence in Policing & Public Safety Program at the University of South Carolina Joseph F. Rice School of Law.

Brandon del Pozo, PhD, MPA, MA, studies public safety and public health at Brown University. Prior to research, he spent 23 years as a police offier, including 19 in the New York City Police Department, where he patrolled Brooklyn, the Bronx, and Manhattan.

Irick A. Geary Jr. is a Ph.D. student in the Department of Criminology and Criminal Justice at the University of South Carolina and a career police commander with a state law enforcement agency.

Marc Olson is a Ph.D. student in Criminology & Criminal Justice at the University of South Carolina. His research focuses on policing and applying quantitative methods to examine criminal justice policy and practice.

Geoffrey Alpert, PhD, is a Professor of Criminology and Criminal Justice at the University of South Carolina, with a concurrent appointment at Griffith University in Brisbane, Australia.