Forensic Science International 315 (2020) 110433

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Juror appraisals of forensic evidence: Effects of blind proficiency and cross-examination

William E. Crozier[a,*], Jeff Kukucka[b], Brandon L. Garrett[a]

[a] Duke University, United States
[b] Towson University, United States

ARTICLE INFO

Article history:
Received 11 May 2020
Received in revised form 1 July 2020
Accepted 22 July 2020
Available online 24 July 2020

Keywords:
Forensic science
Proficiency testing
Expert testimony
Cross-examination
Jury decision-making

ABSTRACT

Forensic testimony plays a crucial role in many criminal cases, with requests to crime laboratories steadily increasing. As part of efforts to improve the reliability of forensic evidence, scientific and policy groups increasingly recommend routine and blind proficiency tests of practitioners. What is not known is how doing so affects how lay jurors assess testimony by forensic practitioners in court. In Study 1, we recruited 1398 lay participants, recruited online using Qualtrics to create a sample representative of the U.S. population with respect to age, gender, income, race/ethnicity, and geographic region. Each read a mock criminal trial transcript in which a forensic examiner presented the central evidence. The low-proficiency forensic examiner elicited a lower conviction rate and less favorable impressions than the control, an examiner for which no proficiency information was disclosed. However, the high-proficiency examiner did not correspondingly elicit a higher conviction rate or more favorable impressions than the control. In Study 2, 1420 participants, similarly recruited using Qualtrics, received the same testimony, but for some conditions the examiner was cross-examined by a defense attorney. We find cross-examination significantly reduced guilty votes and examiner ratings for low-proficiency examiners. These results suggest that disclosing results of blind proficiency testing can inform jury decision-making, and further, that defense lawyering can make proficiency information particularly salient at a criminal trial.

Forensic science evidence plays a crucial role in many criminal cases, with requests to crime laboratories steadily increasing over the past two decades [1]. Recently, growing concern over the accuracy of judgments reached by forensic examiners has led to a new emphasis on improving standards and quality controls, as well as improving the underlying research foundation for forensic methods [2]. Because many such forensic methods lack objective measures and/or standards, an examiner's subjective judgment plays an important role in reaching conclusions based on crime-scene evidence, and examiners who differ in their skill and experience can reach different conclusions regarding the same evidence [2]. Accordingly, examiners in many forensic domains regularly complete proficiency tests designed to measure their performance, including as part of the routine accreditation of laboratories [3]. As of 2014, 98 % of all publicly-funded crime laboratories had adopted some form of proficiency testing [4].

One goal of proficiency testing is to assess how accurate a particular examiner's judgments are [5,6]. As it stands, however, the results of proficiency tests are of "limited value" [7 p. 51,2]) insofar as they often do not include realistic or challenging materials [8] and/or are non-blind (i.e., the examiner is aware that they are being tested). In contrast, blind proficiency tests are designed to mitigate observer effects and cognitive biases—i.e., that people tend to behave unrealistically when they know that they are being observed [9] and interpret evidence in line with their expectations ([10]; see Ref. [11], for application in the forensic sciences). Scientific and policy groups such as the American Association for the Advancement of Science [7], the President's Council of Advisors on Science and Technology [2], and the National Commission on Forensic Science [12] have all recommended blind proficiency testing over non-blind testing, largely because blind tests are assumed to yield more informative estimates of examiners' real-world performance.

Although blind testing presents some practical and logistical challenges [13], at least one large U.S. laboratory—the Houston Forensic Science Center—instituted a laboratory-wide blind proficiency testing program in recent years, intended to improve quality control at every stage of the lab's operations [14], as have

* Corresponding author at: Duke University School of Law, 210 Science Drive, Durham, NC 27708-0373, United States.
E-mail address: william.crozier@duke.edu (W.E. Crozier).

http://dx.doi.org/10.1016/j.forsciint.2020.110433
0379-0738/

laboratories in a number of other counties [5]. Proficiency testing is routine for clinical laboratories in the U.S. that, for example, visually examine cell samples for cancer or pre-cancerous conditions [5]. Performance on blind tests can then be used to identify techniques that should be refined or abandoned, or examiners who stand to benefit from feedback on their performance and/or additional training [5].

As blind proficiency testing becomes more widespread, the results of such tests are likely to be reported and the subject of forensic examiners' testimony in court. However, little is known about how such information will impact fact-finders' appraisals of examiners' opinions. Mitchell and Garrett [3] recently found that a fingerprint examiner's proficiency test performance predicted how much weight jurors placed on the testimony. In that study, a mock fingerprint examiner described his historical performance on proficiency tests as either 100 %, 98 %, 92 %, 86 %, or 66 % correct. In turn, mock jurors discounted the strength of the evidence, and were less likely to convict, in a linear fashion corresponding to the examiner's self-reported error rate. Notably, a Control condition in which the examiner made no mention of proficiency testing yielded results similar to those of the 100 % and 98 % proficient conditions, suggesting that jurors assume examiners to be highly proficient unless explicitly told otherwise. However, this experiment leaves much unknown in terms of how proficiency information impacts jurors' appraisals of forensic evidence and decision-making. For instance, Mitchell and Garrett's mock-examiner did not complete *blind* proficiency tests, was always a fingerprint examiner, and was never cross-examined.

Further, Mitchell and Garrett [3] found that laypeople assume fingerprint examiners are highly proficient, but it is possible that laypeople do not make this same assumption for other forensic disciplines, such that the effect of proficiency information may vary between forensic science disciplines and/or be amplified by effective cross-examination. Indeed, Garrett et al. [15] provided laypeople with error-rate information regarding both fingerprint and voice comparison evidence, and they found that this information reduced guilty verdicts in the fingerprint condition, but it did not affect the weight given to voice comparison evidence, which was universally low. However, their study did not manipulate the individual examiner's error rate, but rather presented jurors with general error rates for each domain. Additional work has examined the impact of calling a rebuttal defense expert on participants' judgments of experts, and found defense expert testimony to reduce guilty votes by lay participants (Garrett & Mitchell, in press).

Because blind proficiency testing yields relevant information about the quality of an individual examiner's work, it may affect juror evaluations of examiners via multiple paths. Kaasa et al. [16] suggested a two-part decision-making process, in which jurors must assess the *reliability* and *diagnosticity* of an examiner's testimony. Specifically, the juror must decide if the forensic method is valuable for proving the underlying fact an examiner is proffering (reliability), before assessing the value of the examiner's conclusion that two pieces of evidence originated from the same source (diagnosticity). Put another way, reliability pertains to a more general appraisal of the methods and underlying science, whereas diagnosticity pertains to the specific case and evidence in question. Because blind proficiency testing does not tell us anything about the evidence in a specific case, it would not aid a juror in assessing diagnosticity. It does, however, tell us something generally about the examiner and their methods – that is, whether or not their methods reliably lead to a correct conclusion. Thus, we conceptualize blind proficiency testing as informing jurors as to the *reliability* assessment, ideally sensitizing jurors to information about the examiner and their methods outside the context of the specific case facts.

The current studies extend previous work i.e., Ref. [3] by examining how knowledge of a forensic examiner's performance on blind proficiency tests impacts jurors' appraisals of said examiner's testimony. We examine: (1) whether *blind* proficiency testing impacts jurors' perceptions and decision-making; (2) whether this effect differs between forms of forensic evidence that vary in their scientific validity; and (3) whether cross-examination moderates this effect. To the first point, examiners may worry that jurors who know that they have made errors on proficiency tests may find their testimony less credible and persuasive. On the other hand, jurors may view an examiner as *more* credible if aware that the examiner adheres to best practices and had the opportunity to identify and learn from mistakes.

To the second point, research has been done on a range of forensic techniques, which shows generally how laypeople place strong weight on forensic evidence e.g., Refs. [17–19]. Prior research has found that jurors tend to place strong weight on fingerprint evidence, regardless of how the examiner phrases ultimate conclusions [20]. However, jurors' beliefs about the reliability of other forms of forensic evidence appear to vary between disciplines [15,21,22]. Unlike fingerprint evidence, few studies have been done examining how jurors evaluate bitemark evidence. In one study, Ribeiro et al. [22] observed that participants rated forensic dental comparison as highly accurate and just as accurate as fingerprint analysis—despite the fact that perhaps the most popular of which, bitemark comparison, has been denounced as lacking foundational validity and therefore not suitable for use in court [2,23–25]. Importantly, however, participants in Ribeiro et al. [22] rated dental analysis as more subjective than fingerprint analysis, suggesting that proficiency information may have a greater impact on jurors' perceptions of a bitemark examiner than of a fingerprint examiner. Elsewhere, Koehler [21] found that people estimated bitemark errors to occur once out of one million comparisons – an error rate higher than fingerprints (one in five million), but still relatively uncommon.

Of course, effective cross-examination may make issues of proficiency and/or subjectivity more salient to jurors. Prior work has examined the effect of cross-examination, including on expert and forensic evidence. Garrett and Mitchell [20] found that regardless of whether the fingerprint examiner acknowledged limitations of fingerprint comparison methods during direct examination or cross-examination, that acknowledgment led lay participants to give the fingerprint evidence less weight. Similarly, Lieberman et al. [26] found DNA evidence strongly influenced guilty decisions, even after a cross-examination. Ziemke and Brodsky [27] examined an "inoculation tactic," whether having an expert acknowledge during direct examination he was being paid and often testified for defendants would neutralize a prosecution effort to portray the expert as a "hired gun;" they found no effects of doing so. Thompson and Scurich [28] found mock jurors rated a forensic odontologist as less credible and were less likely to convict when the examiner admitted on cross examination that the bitemark examination rested on his subjective judgment, and that he was exposed to potentially biasing task-irrelevant contextual information, relative to when these issues were not raised. On the other hand, McQuiston-Surret and Saks [29] found that cross-examination only affected jurors' understanding of forensic evidence when the expert's findings were framed as a subjective probability. Similarly, Koehler [30] found cross-examination of a shoe print examiner had no significant effect on the majority of dependent variables (including verdict and probability of guilt), and made only judgments of shoe print evidence seem slightly weaker and less convincing. Thus, some results suggest both inoculation against cross-examination, and conducting blind proficiency testing, might improve juror ratings of the evidence, while others suggestion no effects. The cross-examination in those

studies, however, did not specifically touch on examiner proficiency — but addresses an problem that blind proficiency reduces.

Finally, as proficiency testing has become quite common and blind proficiency testing has become more common, practitioners may have practical concerns over whether and how to disclose test results in the courtroom. Based on our conversations with practitioners, they may fear that even a single error on a proficiency test might destroy an examiner's credibility in the eyes of the jury, or that information about such testing might induce general skepticism of their entire domain. Or they may fear that laypeople will place undue weight on proficiency information and not focus on the work done to examine evidence in a particular case. Existing research does not shed light on any of those concerns that practitioners would understandably share.

1. Study 1

In Study 1, participants read a case summary and testimony from a forensic examiner in which we varied information about the examiner's blind proficiency performance, allowing us to test for effects on mock jurors' verdicts, evaluations of the evidence, and perceptions of the examiner. We compared fingerprint examiners to bitemark examiners, reasoning that proficiency information might have a lesser effect on a more highly-regarded form of evidence (fingerprints). We also tested the extent to which participants held baseline assumptions about an examiner's proficiency by comparing ratings of high- and low-proficiency examiners against examiners who either provided no proficiency information or claimed high proficiency without offering any proof. Because many examiners claim to be infallible [31], we presume it is not unusual for real world jurors to hear such a claim, yet it is unclear whether jurors accept such a claim at face value.

Study 1 was pre-registered, including the hypotheses, a priori power analysis, materials, and analysis plan. All of these components, as well as the data, are available on the Open Science Framework, https://osf.io/vq3xa/?view_only=9922040f648046febc38650765bfeb95 [link anonymized for peer review].

2. Study 1 method

2.1. Participants and design

Participants ($N = 1420$) were recruited via Qualtrics and completed the study online. Each participant was randomly assigned to one of eight cells in a 2 (Evidence: Bitemark vs. Fingerprint) X 4 (Proficiency: Unknown, Low, High, or High-Unproven) between-subjects design. To ensure that all participants were eligible to serve on a jury in the United States, we later excluded data from 22 individuals (1.5 %) who self-reported a prior felony conviction or pending felony charge, leaving a final sample of $N = 1398$. This sample size was determined a priori using G*Power [32], yielding 90 % power to detect small differences in verdicts ($OR = 1.25$) and small-to-medium differences in a continuous likelihood of guilt measure (Cohen's $d = 0.35$) between any two groups.

All participants were current U.S. citizens. Our sample was stratified to be representative of the U.S. population, with a slight female majority (50.5 %) and a mean age of 47.48 ($SD = 16.58$; range = 18–90). Most participants self-identified as White (61.7 %), with others self-identifying as Hispanic (16.8 %), Black (13.0 %), Asian (5.2 %), Native American (1.5 %), or other race (1.8 %). The sample included at least one resident from 47 of the 50 U.S. states, with the four census regions (i.e., Northeast, South, Midwest, and West) proportionately represented (i.e., 19.2 %, 31.1 %, 24.7 %, and 25.0 %, respectively). The modal participant had completed some college (34.3 %) and self-reported a gross household income of $100,000–$149,999 (14.6 %), with 52.5 % of participants making $59,000 or less a year.

2.2. Procedure

After providing consent, participants provided demographic information and were randomly assigned to one of eight conditions. All participants then read a brief description of a robbery in which the only available evidence was either a bitemark on the victim's arm or a fingerprint on the robber's gun, followed by the testimony of either a bitemark or fingerprint examiner who opined that the forensic evidence implicated the defendant. For some participants, the examiner explained blind proficiency testing and reported that he had either made no errors (*High Proficiency* condition) or six errors (*Low Proficiency* condition) on approximately 20 blind proficiency tests in the past year. For others, the examiner did not explain proficiency testing and either claimed to be highly proficient without offering any proof (*High-Unproven* condition) or did not characterize his own proficiency (*Control* condition). Finally, participants rendered a verdict, estimated the likelihood of the defendant's guilt, provided opinions of the examiner and the evidence, and completed a manipulation check.

2.3. Materials

2.3.1. Case summary

Participants read one of two versions of a 314-word case summary in which a man robs a convenience store and police identify a suspect who is now on trial for armed robbery. Both versions explained that the culprit, who wore a mask and could not be identified by any witnesses, walked into a convenience store, brandished a gun, and demanded the contents of the cash register. As the cashier was pulling out money from the register, she pressed a hidden button to activate an alarm that called the police. When the cashier did not immediately hand over the cash, the robber grabbed and bit her arm so that she dropped the bills. Grabbing the money, the robber rushed out of the store and dropped his gun when his hand caught on the door.

The two versions of the case summary differed with respect to their description of the forensic evidence. Participants in the Fingerprint condition were told that the bitemark on the cashier's arm was not of sufficient quality to be analyzed, but the robber left a fingerprint on the gun that was suitable for analysis. Participants in the Bitemark condition were told the opposite—that the bitemark was suitable for analysis, but the fingerprint on the gun was insufficient to analyze.

The detectives used the cashier's description of the perpetrator's clothing to arrest a man (i.e., the defendant), who was spotted wearing similar clothing in the vicinity of the store shortly after the robbery. The prosecution's case therefore hinged on a single piece of forensic evidence: a forensic examiner's opinion of either a bitemark or fingerprint from the crime scene.

2.3.2. Examiner testimony

Participants read a mock trial transcript (670–891 words) from either a bitemark or fingerprint examiner (corresponding with experimental condition) who had reviewed the evidence in this case. First, the examiner described their credentials. They had worked in the state crime lab as a latent fingerprint specialist for the last 12 years. Prior to that, they had worked at the FBI as a fingerprint examiner for two years: the first in training, and the second filing, searching, and retrieving fingerprint evidence. The examiner also notes that they regularly attend conferences, training seminars, and instructional classes.

Table 1
Effects of Proficiency on Guilt Judgments and Perceptions of Examiner and Evidence (Study 1).

	Unknown	Low	High	High-unsourced
Guilty verdicts (%)	75.2_a	65.1_b	77.7_a	79.2_a
Likelihood of commission (0–100)	79.36_{ab} (23.58)	75.23_a (24.62)	80.97_b (24.36)	81.44_b (22.65)
Examiner convincing (1–7)	6.24_a (1.23)	5.97_b (1.32)	6.41_a (1.12)	6.39_a (1.07)
Examiner competence (1–7)	6.48_a (0.98)	6.08_b (1.13)	6.58_a (0.88)	6.53_a (0.89)
Examiner skill (1–7)	6.56_a (0.88)	6.18_b (1.09)	6.62_a (0.82)	6.60_a (0.82)
Examiner confidence (1–7)	6.58_a (0.83)	6.31_b (1.06)	6.66_a (0.73)	6.63_a (0.74)
The forensic evidence presented by the examiner was persuasive. (1–7)	6.16_a (1.33)	5.89_b (1.41)	6.25_a (1.32)	6.30_a (1.16)
The forensic analysis was based on good scientific principles. (1–7)	6.26 (1.23)	6.18 (1.16)	6.31 (1.27)	6.29 (1.16)
The forensic analysis followed a clearly-defined and standard procedure. (1–7)	6.44 (1.07)	6.41 (1.03)	6.54 (1.00)	6.51 (0.95)

Note: Values not sharing a common subscript differ at $p < 0.05$.

The examiner then explained they had analyzed the relevant evidence and concluded that the evidence implicated the defendant (i.e., that the bitemark matched the defendant's teeth or the latent print matched the defendant's finger). After stating this conclusion, the examiner responded to one of four possible blocks of questions about blind proficiency testing and their recent performance on these tests (i.e., Proficiency manipulation):

1. The examiner explained the procedure and benefits of blind proficiency testing, and he reported that he had previously made six errors on proficiency tests, and had been given over twenty tests in the last year (*Low* proficiency condition).
2. The examiner explained open proficiency testing, why blind proficiency testing is better, and reported that he had never made an error on a proficiency test, and had been given over twenty tests in the last year (*High* proficiency condition).
3. The examiner did not explain proficiency testing. Instead, he stated that he regularly participates in training programs and discusses research and methods with his colleagues, and he is not aware of having ever made any errors in his work (*High-Unproven* condition).
4. The examiner did not explain proficiency testing, nor did he characterize his own accuracy or performance (*Control* condition).

2.3.3. Dependent measures

After reading the examiner's testimony, participants rendered a dichotomous verdict and estimated the likelihood that the defendant had committed the crime (ELOC; 0–100%). Participants also rated the examiner's convincingness, competency, skill, and confidence—each on a scale from 1 (not at all) to 7 (extremely). They also rated the degree to which the forensic analysis was persuasive, based on good scientific principles, and followed a clearly-defined and standard procedure—each on a scale from 1 (completely disagree) to 7 (completely agree).

Next, participants rated the degree to which they found the examiner's experience, qualifications, explanation of the evidence and analysis in general, explanation of the evidence and analysis in this case, and ultimate conclusion in this case—each on a 1 (less convincing) to 7 (more convincing) scale. Using this same scale, participants in the Low and High Proficiency conditions also rated the persuasiveness of the examiner's explanation of proficiency testing.

2.3.4. Manipulation and attention checks

Participants responded to an instructional manipulation check that was embedded among the dependent variables [33], and designed to identify participants who were not reading all materials. Further, because of the response needed to pass, it is extremely improbable they would do so by chance:

In criminal trials, who presents information can affect a jury's perception of that information. Usually the prosecution goes first and presents their evidence – including the witness you heard here. Then, the defense has a chance to present their case, including their own witnesses and evidence that they think may cast reasonable doubt. Ignore the rest of this question, it is a test to make sure you are paying attention. Below, choose the other option and put a plus sign (+) in the box. The prosecution and defense go back and forth in each phase of the trial, including opening statements, primary case, and closing statement.

In this case, information about the expert's past experience might be relevant. Who do you think would be the best person to ask the expert about his past?

Participants then selected from one of the following options: The prosecution, the defense, the judge, an independent lawyer not affiliated with either side, Other (please specify, with a text box).

After providing all dependent measures, participants also completed a multiple-choice comprehension test in which we asked them to identify the charge against the defendant (armed robbery), the source of the forensic evidence (robber's gun or cashier's arm, depending on Evidence condition), and which agencies the examiner had previously worked for (both the FBI and state crime lab). Participants who failed the instructional manipulation check ($n = 4502$) or answered any of the comprehension test items incorrectly ($n = 463$) were excluded.[1]

2.4. Hypotheses

1. Participants who read the testimony of a fingerprint examiner will generate more guilty verdicts, higher ELOC estimates, and more favorable impressions of the examiner and evidence compared to those who read the testimony of a bitemark examiner.
2. Participants who read the testimony of a high proficiency examiner will generate the most guilty verdicts, highest ELOC estimates, and most favorable impressions of the examiner and evidence. Conversely, those who read the testimony of a low proficiency examiner will generate the fewest guilty verdicts, lowest ELOC estimates, and least favorable impressions of the examiner and evidence.
3. The effect of proficiency information will depend on evidence type, such that proficiency information will have a stronger effect on judgments and perceptions in the bitemark condition than in the fingerprint condition.

[1] While these exclusion rates are extremely high, our exclusion criteria were determined *a priori* to ensure that our data would yield accurate estimates of the hypothesized effects (see Ref. [33])—especially in light of growing concern over the validity of online data (see, e.g. Refs. [52,53]). To the extent that these exclusions limit generalizability, it does so only for individuals who were not paying attention, which would ideally not be the case for actual jurors in real-world trials.

3. Study 1 results

3.1. Verdicts

Overall, 74.2 % of participants returned a guilty verdict. A logistic regression using Evidence, Proficiency, and their interaction as predictors of verdict revealed that Evidence did not predict verdict, Wald $\chi^2(1) = 0.49$, $p = 0.483$, $OR = 1.19$ [95 % CI: 0.73, 1.92], such that participants in the Bitemark (72.1 %) and Fingerprint (76.4 %) conditions were similarly likely to vote guilty.

Proficiency did predict verdict, Wald $\chi^2(3) = 16.25$, $p = 0.001$. As shown in Table 1, participants in the Low Proficiency condition were less likely to vote guilty than those in the Unknown, High, and High-Unproven conditions—which did not differ from each other. The Evidence X Proficiency interaction was not significant, Wald $\chi^2(3) = 1.04$, $p = 0.791$.

3.2. Likelihood of commission

Participants generally thought it likely that the defendant had committed the crime (overall $M = 79.23$, $SD = 23.92$). A 2 (Evidence) X 4 (Proficiency) ANOVA on ELOC ratings revealed a main effect of Evidence, $F(11,390) = 4.04$, $p = 0.045$, $d = 0.21$ [95 % CI: 0.10, 0.31], such that ELOC ratings were higher in the Fingerprint condition ($M = 80.51$, $SD = 22.43$) than in the Bitemark condition ($M = 77.94$, $SD = 25.29$).

A main effect of Proficiency also emerged, $F(31,390) = 4.90$, $p = 0.002$, $f = 0.10$. As shown in Table 1, participants in the Low Proficiency condition believed less strongly in the defendant's guilt than those in the High or High-Unproven conditions—none of which differed from the Control condition. No Evidence X Proficiency interaction was found, $F(31,390) = 0.31$, $p = 0.818$, $f = 0.03$.

3.3. Perceptions of the examiner

Across conditions, participants rated the examiner as highly convincing ($M = 6.25$ out of 7, $SD = 1.20$), competent ($M = 6.42$, $SD = 1.00$), skilled ($M = 6.49$, $SD = 0.93$), and confident ($M = 6.54$, $SD = 0.86$).

A 2 (Evidence) X 4 (Proficiency) MANOVA on these ratings revealed a multivariate effect of Proficiency, $F(124,167) = 5.76$, $p < 0.001$, $\eta^2_p = 0.02$, with significant univariate effects on all four items, all $ps < 0.001$, $fs \geq 0.15$. For each of these items, participants in the Low Proficiency condition rated the examiner less favorably than the other three conditions, which did not differ from each other (see Table 1). Neither the effect of Evidence, $F(41,387) = 1.43$, $p = .222$, $\eta^2_p = .00$, nor the interaction, $F(124,167) = 0.87$, $p = 0.578$, $\eta^2_p = 0.00$, was significant.

3.4. Perceptions of the evidence

Overall, participants tended to believe that the forensic evidence was persuasive (overall $M = 6.15$, $SD = 1.32$) and that its analysis was based on good scientific principles ($M = 6.26$, $SD = 1.21$) and followed a clearly-defined procedure ($M = 6.48$, $SD = 1.02$).

A 2 (Evidence) X 4 (Proficiency) MANOVA on these ratings revealed a multivariate effect of Evidence, $F(31,388) = 7.46$, $p < 0.001$, $\eta^2_p = 0.02$. Compared to the Bitemark condition, participants in the Fingerprint condition found the evidence more persuasive ($Ms = 6.05$ & 6.24, $SDs = 1.39$ & 1.23, respectively, $F(11,390) = 12.57$, $p = 0.007$, $d = 0.15$ [95 % CI: 0.04, 0.25], and felt more strongly that its analysis was based on good scientific principles ($Ms = 6.13$ & 6.39, $SDs = 1.33$ & 1.05, respectively), $F(11,390) = 23.33$, $p < 0.001$, $d = 0.22$ [95 % CI: 0.11, 0.32]. Evidence did not affect ratings of the analytic procedure, $F(11,390) = 1.48$, $p = 0.225$, $d = 0.07$ [95 % CI: -0.04, 0.17].

A multivariate effect of Proficiency also emerged, $F(94,170) = 3.53$, $p < 0.001$, $\eta^2_p = 0.01$. Proficiency affected ratings of persuasiveness, $F(31,390) = 11.85$, $p < 0.001$, $f = 0.12$, such that the Low Proficiency examiner was rated as less persuasive than the other three conditions, which did not differ (Table 1). However, Proficiency did not affect perceptions of scientific validity, $F(31,390) = 0.75$, $p = 0.525$, $f = 0.04$, or the analytic procedure, $F(31,390) = 1.19$, $p = 0.311$, $f = 0.05$. No multivariate interaction was found, $F(94,170) = 1.44$, $p = 0.166$, $\eta^2_p = 0.00$.

3.5. Elements of examiner testimony

A 2 (Evidence) X 4 (Proficiency) X 5 (Element) mixed ANOVA was performed on participants' ratings of the degree to which they found five distinct elements of the examiner's testimony (i.e., his experience, qualifications, explanation of the analysis in general, explanation of the analysis in this case, and conclusion in this case) to be convincing. A main effect of Element was found, $F(45,560) = 48.71$, $p < 0.001$, $f = 0.19$. Post hoc Bonferroni comparisons showed that participants rated the examiner's experience ($M = 6.38$, $SD = 1.02$) and qualifications ($M = 6.36$, $SD = 1.01$) as more convincing than his explanation of the analysis in general ($M = 6.24$, $SD = 1.15$), which they in turn rated as more convincing than his explanation of the analysis ($M = 6.15$, $SD = 1.20$) and conclusion in this case ($M = 6.12$, $SD = 1.24$).

This effect was qualified by an Evidence X Element interaction, $F(45,560) = 3.05$, $p = 0.016$, $f = 0.05$. Simple effects tests indicated that participants in the Bitemark and Fingerprint conditions were similarly convinced by the examiner's experience, $t(1396) = 1.02$, $p = 0.307$, $d = 0.05$ [95 % CI: 0.06, 015], and qualifications, $t(1396) = 1.19$, $p = 0.233$, $d = 0.07$ [95 % CI: −0.04, 0.17], but those in the Fingerprint condition were more convinced by the examiner's explanations of the analysis in general, $t(1396) = 2.75$, $p = 0.006$, $d = 0.15$ [95 % CI: 0.04, 0.25], and in this case, $t(1396) = 2.32$, $p = 0.021$, $d = 0.13$ [95 % CI: 0.02, 0.23], and by his conclusion in this case, $t(1396) = 2.86$, $p = 0.004$, $d = 0.15$ [95 % CI: 0.05, 0.26]. Neither a Proficiency X Element, $F(125,560) = 0.60$, $p = 0.843$, $f = 0.04$, nor a three-way interaction, $F(125,560) = 0.66$, $p = 0.788$, $f = 0.04$, was found.

3.6. Self-reported impact of proficiency information

Overall, participants in the Low and High Proficiency conditions felt that the examiner's explanation of proficiency testing made his testimony more convincing ($M = 5.34$, $SD = 1.78$). A 2 (Evidence) X 2 (Proficiency: Low vs. High) ANOVA on these ratings revealed an effect of Proficiency, $F(1693) = 96.87$, $p < 0.001$, $d = 0.57$ [95 % CI: 0.60, 0.91], such that participants in the High condition rated the proficiency information as more convincing ($M = 5.97$, $SD = 1.40$) than those in the Low condition ($M = 4.72$, $SD = 1.89$). Neither the effect of Evidence, $F(1693) = 1.70$, $p = 0.192$, $d = 0.09$ [95 % CI: −0.06, 0.24], nor the Evidence X Proficiency interaction, $F(1693) = 0.23$, $p = 0.634$, $f = 0.02$, was significant.

4. Study 1 discussion

In Study 1, the low-proficiency forensic examiner elicited a lower conviction rate and less favorable impressions than the control (i.e., unknown proficiency) examiner. However, the high-proficiency examiner did not correspondingly elicit a higher conviction rate or more favorable impressions than the control examiner. Moreover, and consistent with Mitchell and Garrett [3], jurors viewed the examiner who claimed high proficiency without any proof no differently than the control or high-proficiency

examiners, which suggests that jurors assume forensic examiners to be highly accurate unless explicitly informed otherwise.

While the results of Study 1 are informative, its ecological validity is limited by the absence of a cross-examination component. In Study 1, jurors' ratings of the forensic examiner's skill and persuasiveness were consistently near-ceiling—perhaps because they did not also read a cross-examination in which an attorney highlighted the examiner's weaknesses or limitations as they presumably would in a real-world trial. As such, we reasoned that cross-examination may amplify the effect of proficiency information by reiterating that information and/or explaining limitations of the analysis that jurors would not otherwise intuit.

5. Study 2

Study 2 was designed as a partial replication and extension of Study 1. First, all participants read the case summary from Study 1 and the testimony of either the low-, high-, or unknown (i.e., control) proficiency fingerprint examiner. Then, some participants also read a cross-examination of the fingerprint examiner, which reiterated his performance on blind proficiency tests and/or highlighted the subjectivity inherent to his analysis. Lastly, participants rendered a verdict and provided the same dependent measures collected in Study 1.

As with Study 1, Study 2 was fully pre-registered. All components of the pre-registration, experiment, data, and analyses can be found on OSF at https://osf.io/evy5a/?view_only=de570b7362054ecb859c924b77d0adbf.

6. Study 2 method

6.1. Participants and design

Participants ($N = 1412$) were recruited via Qualtrics and completed the study online. Each participant was randomly assigned to one of seven cells in a 2 (Proficiency: Low vs. High) X 3 (Cross-examination: None, Testing, or Testing + Subjectivity) + 1 (Control: Unknown Proficiency and No Cross-examination) between-subjects design. All participants were eligible to serve on a jury in the United States. As with Study 1, we preregistered this sample size as sufficient to detect small-medium differences in ELOC and Likert Scale ratings between any two groups ($d = 0.35$) and an odds ratio of 1.25 for differences in guilty verdicts with 90 % power.

As in Study 1, our sample was stratified to be representative of the U.S. population, with a slight male majority (50.4 %) and a mean age of 46.83 ($SD = 17.22$; range = 18–91). Most participants self-identified as White (63.7 %), with others self-identifying as Hispanic (16.3 %), Black (10.9 %), Asian (5.2 %), Native American (1.1 %), or other race (2.8 %). The sample included at least one resident from 49 of the 50 U.S. states, with the four census regions proportionately represented (i.e., 16.6 % Northeast, 38.2 % South, 21.9 % Midwest, and 23.3 % West). The modal participant had completed some college (39.9 %) and self-reported a gross household income of $30,000–$39,999 (12.1 %) and 60.5 % making $59,000 or less a year.

6.2. Procedure

The procedure of Study 2 was similar to Study 1, with the addition of a cross-examination of the examiner. Participants first read a set of case facts describing a crime which included only fingerprint evidence, as we eliminated the bitemark manipulation. Participants then read a transcript of a forensic analyst's testimony and subsequent cross-examination. They then provided a verdict, ELOC rating, and views of the examiner and the evidence, as in Study 1.

6.3. Materials

6.3.1. Case summary

The case summary was identical to the summary used in the Fingerprint conditions of Study 1 (i.e., the only evidence against the defendant was the opinion of a fingerprint examiner).

6.3.2. Examiner testimony

By random assignment, participants read the testimony of either the Low-Proficiency, High-Proficiency, or Control (i.e., unknown proficiency) examiner—each of which was identical to Study 1.

6.3.3. Cross-examination

Some participants also read a cross-examination of the fingerprint examiner. To be exact, participants in the *Testing* cross-examination condition read a cross-examination in which the defense attorney emphasized that blind proficiency tests are meant to ensure examiners are doing their work properly, and reiterated the number of errors that the examiner had made on these tests. Participants in the *Testing + Subjectivity* cross-examination condition read a cross-examination that included the aforementioned elements as well as an exchange in which the examiner stated that fingerprint examiners often deal with incomplete, distorted, or smudged prints, and do not use objective measures in their analyses, such that errors sometimes occur and two different examiners could arrive at different opinions about the same fingerprints. Participants in the *None* and *Control* conditions did not read any cross-examination of the forensic examiner.

6.3.4. Dependent measures

The dependent measures in Study 2 were identical to those collected in Study 1.

6.3.5. Instructional manipulation and attention checks

As in Study 1, all participants who responded incorrectly to the instructional manipulation check ($n = 4071$) or attention checks (1590) were excluded.

6.4. Hypotheses

1. We predict an ANOVA main effect of Proficiency, such that high proficiency examiner will result in more guilty verdicts, higher ELOC, higher examiner ratings, and higher evidence ratings, than low proficiency examiner. The *No Cross* condition will not be significantly different than either high or low proficiency (as assessed by t-tests)
2. We predict an ANOVA main effect of Cross-Examination type, such that participants in the *No Cross* category will provide more guilty verdicts, higher ELOC, higher examiner ratings, and higher evidence ratings than the *Testing Cross* and *Testing + Subjectivity Cross* conditions. We also predict that *Testing + Subjectivity Cross* will be significantly lower than *Testing Cross* on those DVs. Although Thompson and Scurich [28] did not find an additive effect of cross examining on bias and subjectivity, our results from Study 1 demonstrated low proficiency test performance significantly lowered verdicts and ELOC. Thus, if the cross-examination has the effect we predict, it is possible that we will see a main effect of Cross-Examination type driven primarily by differences in the Low proficiency condition (that is, qualified by an interaction term as specified below). Participants in the hanging control will be significantly higher on DVs than participants in the *Testing Cross* and *Testing + Subjectivity Cross* (as assessed by t-tests).
3. We predict a significant interaction between Proficiency and Cross-Examination, such that the gap between *High Proficiency*

Table 2
Effects of Proficiency and Cross-examination on Guilt Judgments and Perceptions of Examiner and Evidence (Study 2).

		Low Proficiency			High Proficiency		
	Control	None	Test	Test/Subj	None	Test	Test/Subj
Guilty verdicts (%)	75.1_{ab}	70.8_{ab}	51.3_{c}	37.0_{d}	85.5_{e}	78.6_{be}	68.0_{a}
Likelihood of commission (0–100)	78.22_{a} (25.64)	77.33_{a} (26.32)	66.69_{b} (28.08)	63.11_{b} (27.79)	86.33_{c} (19.51)	80.70_{ac} (22.81)	75.44_{a} (26.36)
Examiner convincingness (1–7)	6.35_{ab} (1.08)	6.10_{a} (1.31)	5.65_{c} (1.47)	5.38_{c} (1.44)	6.66_{b} (0.87)	6.39_{ab} (1.03)	6.31_{a} (0.96)
Examiner competence (1–7)	6.54_{ab} (0.79)	6.14_{c} (1.14)	5.78_{d} (1.24)	5.68_{d} (1.20)	6.78_{a} (0.64)	6.49_{ab} (1.00)	6.43_{bc} (0.94)
Examiner skill (1–7)	6.59_{ab} (0.78)	6.22_{b} (1.14)	5.80_{c} (1.19)	5.78_{c} (1.18)	6.80_{a} (0.62)	6.52_{b} (0.85)	6.54_{ab} (0.74)
Examiner confidence (1–7)	6.55_{ab} (0.85)	6.38_{b} (1.09)	6.02_{c} (1.16)	5.76_{c} (1.20)	6.80_{a} (0.54)	6.56_{ab} (0.87)	6.49_{b} (0.79)
The forensic evidence presented by the examiner was persuasive. (1–7)	6.33_{ab} (1.16)	5.92_{c} (1.45)	5.34_{d} (1.69)	5.15_{d} (1.52)	6.55_{a} (1.04)	6.19_{abc} (1.20)	6.04_{bc} (1.32)
The forensic analysis was based on good scientific principles. (1–7)	6.41_{ab} (1.11)	6.24_{abc} (1.26)	5.91_{c} (1.39)	5.41_{d} (1.44)	6.58_{a} (1.10)	6.38_{ab} (1.01)	6.11_{bc} (1.19)
The forensic analysis followed a clearly-defined and standard procedure. (1–7)	6.56_{a} (0.88)	6.47_{abc} (1.07)	6.23_{bc} (1.15)	5.57_{d} (1.54)	6.74_{a} (0.79)	6.55_{ab} (0.91)	6.19_{c} (1.27)

Note: None = No cross-examination. Test = Testing cross-examination. Test/Subj = Testing/Subjectivity cross-examination. Values not sharing a common subscript differ at $p < 0.05$.

and *Low proficiency* (more guilty verdicts, higher ELOC, higher examiner ratings, and higher evidence ratings in High Proficiency) will be greater for the *Testing Cross* and *Testing + Subjectivity Cross* conditions than in the *No Cross* condition.

7. Study 2 results

7.1. Verdicts

Overall, 66.7 % of participants returned a guilty verdict. A logistic regression using Proficiency, Cross-examination, and their interaction as predictors of verdict revealed that Proficiency predicted verdict, Wald $\chi^2(1) = 12.30$, $p < 0.001$, $OR = 2.43$ [95 % CI: 1.48, 4.00], such that participants in the *High Proficiency* condition voted guilty more often (76.6 %) than those in the *Low Proficiency* condition (53.1 %), replicating results from Study 1.

Cross-examination also predicted verdict, Wald $\chi^2(2) = 44.51$, $p < 0.001$, such that guilty verdicts were most frequent when the examiner was not cross-examined (78.1 %), less frequent for *Testing Cross* participants (65.0 %), and even less frequent for *Testing + Subjectivity Cross* participants (46.4 %). The Proficiency X Cross-examination interaction was not significant, Wald $\chi^2(2) = 1.38$, $p = 0.501$.

As shown in Table 2, relative to the control condition where proficiency was unknown and the examiner was not cross-examined (75.1 %), the *High Proficiency* examiner increased guilty verdicts, but the *Low Proficiency* examiner did not affect verdicts — a trend we did not find in Study 1. Both types of cross-examination (i.e., *Testing Cross* and *Testing + Subjectivity Cross*) lowered the frequency of guilty verdicts in the *Low Proficiency* condition, but only the *Testing + Subjectivity Cross* lowered the frequency of guilty verdicts in the *High Proficiency* condition.

7.2. Likelihood of commission

Participants generally thought it likely that the defendant had committed the crime (overall $M = 75.44$, $SD = 26.38$). A 2 (Proficiency) X 3 (Cross-examination) ANOVA on ELOC ratings revealed a main effect of Proficiency, $F(11,130) = 50.81$, $p < 0.001$, $d = 0.45$ [95 % CI: 0.33, 0.57], such that ELOC ratings were higher in the *High Proficiency* condition ($M = 80.76$, $SD = 23.26$) than in the *Low Proficiency* condition ($M = 69.07$, $SD = 28.02$).

Cross-examination also affected ELOC ratings, $F(21,130) = 28.08$, $p < 0.001$, $f = 0.22$, such that participants were most confident in the defendant's guilt when the examiner was not cross-examined ($M = 81.81$, $SD = 23.59$), less confident for the *Testing Cross* ($M = 73.73$, $SD = 26.49$), and even less confident in the *Testing + Subjectivity Cross condition* ($M = 66.90$, $SD = 27.65$). The Proficiency X Cross-examination interaction was not significant, $F(21,130) = 1.17$, $p = 0.311$, $f = 0.05$.

Echoing the pattern for verdicts, the High Proficiency examiner—but not the Low Proficiency examiner—increased ELOC ratings relative to the control condition. Both types of cross-examination again lowered ELOC ratings in the *Low Proficiency* condition, but only the *Testing + Subjectivity Cross* lowered ELOC ratings in the *High Proficiency* condition (Table 2).

7.3. Perceptions of the examiner

Across conditions, participants rated the examiner as highly convincing ($M = 6.06$, $SD = 1.30$), competent ($M = 6.20$, $SD = 1.12$), skilled ($M = 6.25$, $SD = 1.07$), and confident ($M = 6.32$, $SD = 1.04$). A 2 (Proficiency) X 3 (Cross-examination) MANOVA on these ratings revealed a multivariate effect of Proficiency, $F(41,127) = 35.13$, $p < 0.001$, $\eta^2_p = 0.11$. Compared to the *Low Proficiency* examiner, participants rated the *High Proficiency* examiner as more convincing ($Ms = 5.71$ & 6.45, $SDs = 1.44$ & 0.98, respectively), $F(11,130) = 93.41$, $p < 0.001$, $d = 0.60$ [95 % CI: 0.48, 0.71], more competent ($Ms = 5.87$ & 6.57, $SDs = 1.21$ & 0.89), $F(11,130) = 114.57$, $p < 0.001$, $d = 0.66$ [95 % CI: 0.54, 0.77], more skilled ($Ms = 5.94$ & 6.61, $SDs = 1.19$ & 0.77), $F(11,130) = 122.94$, $p < 0.001$, $d = 0.66$ [95 % CI: 0.54, 0.78], and more confident ($Ms = 6.05$ & 6.61, $SDs = 1.18$ & 0.77), $F(11,130) = 81.77$, $p < 0.001$, $d = 0.56$ [95 % CI: 0.44, 0.68].

A multivariate effect of Cross-examination also emerged, $F(82,256) = 8.04$, $p < 0.001$, $\eta^2_p = 0.03$, with univariate effects on all four items, all $ps < 0.001$, $fs \geq 0.17$. As shown in Table 3, the *Testing Cross* decreased ratings of the examiner along all four dimensions, and the addition of a Subjectivity component decreased ratings of convincingness and confidence even further. The multivariate interaction was not significant, $F(82,256) = 0.56$, $p = 0.815$, $\eta^2_p = 0.00$.

Relative to the control condition, the *Low Proficiency* examiner was seen as less competent, but as similarly convincing, skilled, and confident—and either form of cross-examination decreased ratings of the examiner on all four dimensions (Table 2). Perceptions of the *Control* and *High Proficiency* examiners did not differ; *Testing Cross* decreased only ratings of the examiner's skill, whereas *Testing + Subjectivity Cross* decreased ratings of the examiner's convincingness, competence, and confidence.

Table 3
Effect of Cross-examination on Perceptions of the Expert and Evidence (Study 2).

	None	Proficiency	Proficiency/Subjectivity
Examiner convincingness (1–7)	6.38_a (1.14)	6.02_b (1.32)	5.71_c (1.35)
Examiner competence (1–7)	6.46_a (0.98)	6.14_b (1.18)	5.96_b (1.16)
Examiner skill (1–7)	6.51_a (0.96)	6.16_b (1.10)	6.06_b (1.09)
Examiner confidence (1–7)	6.59_a (0.88)	6.29_b (1.05)	6.01_c (1.12)
The forensic evidence presented by the examiner was persuasive. (1–7)	6.23_a (1.30)	5.77_b (1.56)	5.48_c (1.47)
The forensic analysis was based on good scientific principles. (1–7)	6.41_a (1.19)	6.15_b (1.23)	5.64_c (1.38)
The forensic analysis followed a clearly-defined and standard procedure. (1–7)	6.60_a (0.95)	6.39_b (1.04)	5.78_c (1.44)

Note: Values not sharing a common subscript differ at $p < 0.05$.

7.4. Perceptions of the evidence

Participants generally agreed that the forensic evidence was persuasive (overall $M = 5.85$, $SD = 1.48$) and that its analysis was based on good scientific principles ($M = 6.09$, $SD = 1.30$) and followed a clearly-defined procedure ($M = 6.28$, $SD = 1.19$). A 2 (Proficiency) X 3 (Cross-examination) MANOVA on these ratings revealed a multivariate effect of Proficiency, $F(31,128) = 28.58$, $p < 0.001$, $\eta^2_p = 0.07$. Compared to the *Low Proficiency* condition, participants in the *High Proficiency* condition more strongly believed that the forensic evidence was persuasive ($Ms = 5.47$ & 6.27, $SDs = 1.59$ & 1.21, respectively), $F(11,130) = 85.78$, $p < 0.001$, $d = 0.56$ [95 % CI: 0.44, 0.68], that its analysis was based on good scientific principles ($Ms = 5.86$ & 6.36, $SDs = 1.40$ & 1.12), $F(11,130) = 37.62$, $p < 0.001$, $d = 0.39$ [95 % CI: 0.27, 0.51], and that its analysis followed a clearly-defined procedure ($Ms = 6.09$ & 6.50, $SDs = 1.32$ & 0.99), $F(11,130) = 29.32$, $p < 0.001$, $d = 0.35$ [95 % CI: 0.23, 0.47].

A multivariate effect of Cross-examination also emerged, $F(62,258) = 17.10$, $p < .001$, $\eta^2_p = 0.04$, with univariate effects on all three items, all $ps < 0.001$, $fs \geq 0.20$. For each item, the evidence was rated most favorably when the examiner was not cross-examined, less favorably in the *Testing Cross* condition, and even less favorably for the *Testing + Subjectivity Cross* (Table 3). The multivariate interaction was not significant, $F(62,258) = 0.63$, $p = 0.707$, $\eta^2_p = 0.00$.

Relative to the *Control* condition, the *Low Proficiency* examiner's evidence was rated as less persuasive, but his analysis was rated as similarly scientific and clearly-defined; both cross-examinations weakened the evidence's persuasiveness, but only the *Testing + Subjectivity Cross* affected perceptions of the analysis (Table 2). The evidence and analysis of control and *High Proficiency* examiners were perceived no differently, and the *Testing Cross* did not weaken perceptions of the High Proficiency examiner's evidence or analysis, but the *Testing + Subjectivity Cross* weakened both.

7.5. Elements of examiner testimony

A 2 (Proficiency) X 3 (Cross-examination) X 5 (Element) mixed ANOVA was performed on participants' ratings of how convincing they found five elements of the examiner's testimony. A main effect of Element was found, $F(41,127) = 38.48$, $p < 0.001$, $f = 0.37$; overall, participants were most convinced by the examiner's experience ($M = 6.31$, $SD = 1.02$) and qualifications ($M = 6.26$, $SD = 1.08$), less convinced by his general explanation of the analysis ($M = 6.15$, $SD = 1.17$), even less convinced by his explanation of the analysis in this case ($M = 6.03$, $SD = 1.25$), and even less convinced by his conclusion in this case ($M = 5.95$, $SD = 1.27$).

This effect was qualified by significant Proficiency X Element, $F(41,127) = 2.52$, $p = 0.040$, $f = 0.09$, and Cross-examination X Element, $F(82,256) = 5.09$, $p < 0.001$, $f = 0.13$, interactions. For Proficiency, participants in the *Low Proficiency* condition rated all five elements of the examiner's testimony as less convincing compared to those in the *High Proficiency* condition, all $ts > 5.89$, $ps < 0.001$, $ds > 0.34$. For Cross-examination, the *Testing Cross* and *Testing + Subjectivity Cross* weakened the convincingness of the examiner's experience and qualifications to the same degree, but for the other three elements, *Testing Cross* weakened their convincingness and *Testing + Subjectivity Cross* weakened it even further.

Compared to the control condition, the *High Proficiency* examiner was rated as more convincing on four of the five elements (i.e., all except experience), and neither cross-examination weakened the convincingness of his experience or qualifications, but the *Testing + Subjectivity Cross* weakened the convincingness of the other three elements. The *Low Proficiency* and *Control* examiners were rated as similarly convincing on all five elements, but either type of cross-examination weakened the *Low Proficiency* examiner's convincingness on all five elements (see online supplemental material on OSF for full results).

7.6. Self-reported impact of proficiency information

Overall, participants in the six experimental conditions felt that the examiner's explanation of proficiency testing made his testimony more convincing ($M = 5.11$, $SD = 1.89$). A 2 (Proficiency) X 3 (Cross-examination) ANOVA on these ratings revealed a main effect of Proficiency, $F(21,091) = 153.35$, $p < 0.001$, $d = 0.78$ [95 % CI: 0.65, 0.90], such that participants in the *High* condition rated the proficiency information as more convincing ($M = 5.82$, $SD = 1.49$) than those in the *Low* condition ($M = 4.45$, $SD = 1.99$). A main effect of Cross-examination also emerged, $F(21,091) = 26.77$, $p < 0.001$, $f = 0.22$, such that the *No Cross* condition rated the proficiency information as more convincing ($M = 5.58$, $SD = 1.77$) than the *Testing Cross* condition ($M = 5.15$, $SD = 1.86$), who in turn rated it as more convincing than the *Testing + Subjectivity Cross* condition ($M = 4.50$, $SD = 1.92$). No Proficiency X Cross-examination interaction was found, $F(21,091) = 1.46$, $p = 0.233$, $f = 0.05$.

8. Study 2 discussion

In Study 2, we increased the ecological validity of our study by including a cross-examination of the examiner. A competent defense attorney should not let an examiner go unquestioned, particularly if they claim proficiency without evidence or provide evidence to the contrary. Indeed, defense attorneys may be more skeptical of forensic evidence than laypeople [34]. As such, Study 2 participants read a cross-examination conditions that either did or did not reiterate the examiner's proficiency test performance. As predicted, we found that cross-examining an examiner on their proficiency led lay participants to devalue their testimony. Relative to the low proficiency examiner who was not cross-examined, our Testing cross-examination reduced guilty verdicts, perceived likelihood of commission, and all examiner ratings, and the stronger Testing/Subjectivity cross decreased guilty verdicts and estimations of commission even further. Thus, as expected, highlighting the limitations of the low proficiency examiner's

analysis on cross-examination led participants to give less weight to his testimony.

In contrast, the high proficiency examiner was somewhat protected from the debilitating effect of cross-examination. Relative to the high proficiency examiner who was not cross-examined, only the stronger Testing/Subjectivity cross-examination of the high proficiency examiner decreased guilty votes, likelihood of commission ratings, and examiner ratings. Put another way, the weaker Testing did not faze the high proficiency examiner as it did the low proficiency examiner. This pattern suggests an important applied benefit of blind proficiency testing—namely, that examiners who perform well on such tests are to some degree inoculated against cross-examination.

9. General discussion

These two studies shed light on how jurors react to and utilize information about an individual forensic examiner's performance on blind proficiency tests. Together, they produced some findings that were expected, as well as others that were more nuanced or surprising. Below, we review our principal findings and discuss their implications for forensic science practice and expert testimony.

At first glance, the results of Study 1 could be taken as confirmation of examiners' fears: Insofar as lay participants devalued the low-proficiency examiner but did not correspondingly elevate the high-proficiency examiner, examiners may interpret this pattern as evidence that they have little to gain—but potentially much to lose—from engaging in blind proficiency testing. However, a closer look at our data suggests that these fears are overblown. Although lay participants' devaluation of the low-proficiency examiner was statistically significant in Study 1, it was objectively quite small and participants in both studies rated the low-proficiency examiner very favorably (i.e., $Ms \geq 5.97$ and 6.10, respectively, on a 7-point scale) and still voted to convict 65 % and 71 % of the time when the low-proficiency examiner's testimony was the only evidence against the defendant. Thus, while poor performance on proficiency tests was detrimental to examiners' persuasiveness in Study 1, the degree to which it is detrimental may not be practically important, as ratings of their convincingness and skill were nonetheless consistently near ceiling in both studies.

In fact, Study 2 participants viewed the high proficiency examiner more favorably than the control examiner whose proficiency was unknown. To be exact, participants who read the testimony of a high proficiency examiner returned more guilty verdicts and more favorable ratings of the examiner. Thus, whereas the findings of Study 1 implied that laypeople assume examiners to be highly skilled unless informed otherwise, the results of Study 2 demonstrate that performing well on blind proficiency testing can indeed further enhance examiners' persuasiveness. This finding is in line with our original prediction that high-proficiency examiners should elicit more guilty verdicts, stronger perceptions of guilt, and higher ratings of examiner skill and methodological quality. This result is consistent with prior work suggesting that laypeople can discriminate between levels of proficiency [3].

In both studies, proficiency information informed participants' opinions of the testifying examiner's convincingness, skill, competence, and confidence—but it had little effect on their opinion of the domain's underlying methods or evidence. This pattern suggests that lay participants are able to distinguish between the proficiency of an individual examiner and the reliability of that examiner's method as a whole. In other words, participants' overall impression of the domain and method was likely not soured by one low proficiency examiner—and again, both examiner- and domain-focused ratings were universally high, regardless of proficiency and/or cross-examination.

Theoretically, performance on blind proficiency tests could reflect the validity of an examiner's methodology, insofar as examiners who perform well on proficiency tests are likely using sound methods, and vice versa. However, our data suggest that participants did not consider proficiency testing to be indicative of the validity of the underlying methodology, and they instead attributed proficiency testing performance entirely to the examiner's ability rather than to their methods. This result is consistent with work suggesting that lay participants respond to proficiency information, but retain strong views regarding the reliability of forensic methods in general [3]. Further, it is consistent with work showing that laypeople are generally poor at evaluating the quality of scientific evidence, such as psychological expert testimony [35,36].

Further underscoring participants' limited ability to evaluate scientific evidence, Study 1 participants viewed fingerprint and bitemark examiners as equally skilled and convincing and their testimony elicited similar conviction rates. This finding echoes previous studies in which laypeople believed both fingerprint and forensic dental identification to be somewhat subjective [22] and highly accurate ([21], see also Ref. [37]). This finding is quite concerning, given that latent fingerprint identification has an undeniably stronger scientific foundation than bitemark identification. For instance, whereas fingerprint examiners have demonstrated relatively higher levels of accuracy in the few controlled studies conducted to date, forensic bitemark comparison has been found to lack adequate controlled studies and is therefore lacking in foundational validity [2]. Our results therefore suggest that lay participants fail to appreciate the limitations of bitemark identification and may give an undue amount of weight to this form of evidence.

Though amply powered, our studies are limited by the fact that participants did not deliberate in groups, but rather rendered guilty votes individually, and they read written trial materials which were modeled after testimony from actual criminal cases rather than viewing witness testimony *in vivo*. However, previous research has consistently found that trial simulations based on written materials typically produce the same outcomes as video-based trial simulations [38,39], Conversely, we did sample a large population of census-representative and jury eligible adults. While a recent meta-analysis suggests that college-student samples are representative and suitable for mock jury research, our samples are more ecologically valid as they were stratified to represent the U.S. population as a whole [40].

Another limitation is that Study 2 did not include a condition where the examiner was cross-examined, but did not provide proficiency information, which is presumably a common occurrence in real-world trials. As such, we could not test the potential cost of an examiner admitting under cross-examination that they do not engage in proficiency testing at all. In such a situation, we presume that the examiner would offer an explanation for why they do not complete proficiency tests and/or assert that this does not undermine the validity of their opinion. This rebuttal, while realistic, would introduce a new variable into the cross-examination that would introduce a potential confound into our experimental design. Further, in Study 2 we saw that Control examiners were weakened relatively to High proficiency examiners and participants viewed low-proficiency examiners similarly as no-proficiency examiners, suggesting detrimental impact for a cross-examined no-proficiency examiner.

A redirect examination of a forensic examiner is another ecologically-valid way that an examiner could potentially combat the detrimental effects of a low proficiency test performance. Indeed, with an open-ended questioning format, a redirect may be even more effective than cross-examination responses. A number of possible responses come to mind: an examiner could state their

performance is not, in fact, that problematic; that their mistake on a previous case does not indicate they made a mistake here; that they followed procedure in this case so their conclusions are valid; or even that they have learned from their mistakes and would now pass any blind proficiency test without any errors. Redirect is outside the scope of this article, but future research may want to investigate whether redirect decreases the effects we have found here, either through the aforementioned responses or others.

An additional limitation is that we kept a majority of details about the investigation and trial constant across conditions. Perhaps varying these factors would have moderated the impact of proficiency information. For example, strengthening the prosecution's case by introducing other incriminating evidence may have decreased the impact of proficiency information by prompting less scrutiny of the examiner's incriminating judgment—a form of *forensic confirmation bias* [41]. Similarly, the examiner's experience and qualifications may overshadow their performance on proficiency tests. In both studies, our examiner had 12 years of experience, including employ at the FBI — a qualification that likely impresses jurors e.g., Ref. [42]. In line with a dual-process theory of persuasion, such superficial information tends to affect jurors along a peripheral route, which can lead to less scrutiny put on central-route factors, such as an examiner's methods and performance [43]. Should the examiner be less experienced, jurors may rely more on blind proficiency testing information, and widen the gap between high- and low-proficiency examiners. Similarly, the trial was relatively brief, having only the single form of evidence and single witness. It is possible that in longer trials, with more information, blind proficiency information may lose some of its potency. We note, however, that the forensic examiner's testimony is realistic in content and length, based on our review of hundreds of trial transcripts. Blind proficiency affects only the examiner's testimony and thus the assessment of forensic evidence should be temporally contained to the examiner's time on the stand. Thus, any decrease over lengthier trials in the effects we demonstrate here are likely minimal.

Finally, we acknowledge that the high exclusion rates in both Study 1 and 2 may limit generalizability of these results. Specifically, we excluded participants from analyses because they failed to fully read questions and follow instructions (in the case of the instructional manipulation check), and failed to notice or remember several important case details (in the case of the manipulation checks). These exclusions decrease the noise in our data, increasing its reliability, and increase the precision with which we estimate the effects of our independent variables [33]. It does, however, limit our generalizability to jurors who are paying attention in trial. While it may unfortunately be the case that jurors sometimes do not pay attention, we would not expect the manipulations in these study to affect those jurors, precisely because they did not notice them in the same manner that our excluded participants failed to notice them.

These findings raise practical questions for the criminal legal system. We note that judges have often, in the past, not provided defense lawyers with discovery on proficiency testing information [5]. While constitutional due process rights safeguard access to evidence that can be used to impeach a witness at trial, in the past, judges have sometimes concluded that proficiency information is not valuable as a source of impeachment that could undermine the credibility of an expert witness [5]. For defense lawyers to ask questions regarding proficiency information in a cross-examination, they would obviously first have to have access to it. Further, if there is not routine access to proficiency information, crime laboratories can choose to strategically disclose high proficiency scores, and not disclose low proficiency scores. Our results suggest that they would gain real advantages at a criminal trial by doing so. Our results also suggest the importance of evenhanded discovery regarding such information, discovery of all information which does in fact permit meaningful impeachment of an expert.

Cross-examination is just one tool available at a criminal trial to interrogate evidence. Judges may themselves consider, as part of their gatekeeping responsibilities, whether evidence is sufficiently reliable. Proficiency information may play a role in that task, and it has in the past [5]. Future research could examine whether there are other ways that such information can be made salient to lay jurors, including through jury instructions, a defense expert, closing arguments, or through standards for a discipline that any practitioner would be required to follow, to present the limitations of their methods and conclusions in reports and in courtroom testimony.

10. Conclusion

These results suggest that blind proficiency testing information can valuably inform jury decision-making. These findings may support efforts to conduct blind proficiency testing as part of routine quality programs in crime laboratories, and to share such information with prosecution and defense lawyers in reports and testimony. We find that examiners were not "burned" or found severely undermined by jurors, even when their proficiency scores were quite low, and even when subjected to intense cross-examinations making that poor performance especially salient. To the extent that laboratories fear that blind proficiency testing would create unwarranted challenges in presenting expert testimony in court (see Ref. [14]), we suggest that it does not. These studies support the use of blind proficiency testing more routinely, and they suggest that test results can usefully inform trials, and should be incorporated into the evidence that jurors consider at criminal trials.

Funding

This work was partially supported by the Center for Statistics and Applications to Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between the National Institute of Standards and Technology and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California, Irvine, and University of Virginia.

CRediT authorship contribution statement

William E. Crozier: Conceptualization, Methodology, Software, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing. **Jeff Kukucka:** Conceptualization, Methodology, Software, Investigation, Data curation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Brandon L. Garrett:** Methodology, Resources, Writing - original draft, Writing - review & editing, Supervision.

References

[1] B.L. Garrett, The costs and benefits of forensics, Houst. Law Rev. 56 (2020) 593–616.
[2] President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods, (2016) . http://www.crime-scene-investigator.net/PDF/forensic-science-in-criminal-courts-ensuring-scientific-validity-of-feature-comparison-methods.pdf.
[3] G. Mitchell, B.L. Garrett, The impact of proficiency testing information and error aversions on the weight given to fingerprint evidence, Behav. Sci. Law 37 (1) (2019) 195–210, doi:http://dx.doi.org/10.1002/bsl.2402.
[4] US Department of Justice, Census of Publicly Funded Crime Laboratories Retrieved from:, (2014) . https://www.bjs.gov/index.cfm?ty=dcdetail&iid=244.

[5] B.L. Garrett, G. Mitchell, The proficiency of experts, Univ. Law Rev. 66 (2018) 901–960.
[6] J.J. Koehler, Fingerprint error rates and proficiency tests: what they are and why they matter, Hastings Law J. 59 (2008) 1077–1100.
[7] AAAS. American Association for the Advancement of Sciences, Forensic Science Assessments: a Quality and Gap Analysis, Latent Fingerprint Examination, American Association for the Advancement of Science, New York, NY, 2017.
[8] S. Kelley, B.O. Gardner, D.C. Murrie, K.D. Pan, K. Kafadar, How do latent print examiners perceive proficiency testing? An analysis of examiner perceptions, performance, and print quality, Sci. Justice 60 (2) (2020) 120–127.
[9] M.T. Orne, On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications, Am. Psychol. 17 (11) (1962) 776.
[10] R.S. Nickerson, Confirmation bias: a ubiquitous phenomenon in many guises, Rev. Gen. Psychol. 2 (2) (1998) 175–220, doi:http://dx.doi.org/10.1037/1089-2680.2.2.175.
[11] W.C. Thompson, Interpretation: observer effects, in: A. Moenssens, A. Jamieson (Eds.), Wiley Encyclopedia of Forensic Sciences, John Wiley & Sons, Chichester, UK, 2009, pp. 1575–1579.
[12] National Commission on Forensic Science, Views of the Commission: Facilitating Research on Laboratory Performance (Adopted unanimously 13 September 2016) Available at:, (2016) . https://www.justice.gov/ncfs/page/file/909311/download.
[13] Organization of Scientific Area Committees, Draft Guidance on Testing the Performance of Forensic Examiners Retried from:, (2018) . https://www.nist.gov/sites/default/files/documents/2018/05/21/draft_hfc_guidance_document-may_8.pdf.
[14] C. Hundl, M. Neuman, A. Rairden, P. Rearden, P. Stout, Implementation of a blind quality control program in a forensic laboratory, J. Forensic Sci. 65 (3) (2020) 815–822, doi:http://dx.doi.org/10.1111/1556-4029.14259.
[15] B.L. Garrett, W.E. Crozier, R. Grady, Error rates, likelihood ratios, and jury evaluation of forensic evidence, J. Forensic Sci. (2020), doi:http://dx.doi.org/10.1111/1556-4029.14323.
[16] S.O. Kaasa, T. Peterson, E.K. Morris, W.C. Thompson, Statistical inference and forensic evidence: evaluating a bullet lead match, Law Hum. Behav. 31 (5) (2007) 433–447.
[17] J. Schklar, S.S. Diamond, Juror reactions to DNA evidence: errors and expectancies, Law Hum. Behav. 23 (1999) 159–184.
[18] N. Scurich, R. John, Mock jurors' use of error rates in DNA database trawls, Law Hum. Behav. 37 (2013) 424–431.
[19] W.C. Thompson, S.O. Kaasa, T. Peterson, Do jurors give appropriate weight to forensic identification evidence? J. Empir. Leg. Stud. 10 (2013) 359–397.
[20] B.L. Garrett, G. Mitchell, How jurors evaluate fingerprint evidence: the relative importance of match language, method information and error acknowledgement, J. Empir. Leg. Stud. 10 (2013) 484–511.
[21] J.J. Koehler, Forensics or fauxrensics: ascertaining accuracy in the forensic sciences, Ariz. St. LJ 49 (2017) 1369.
[22] G. Ribeiro, J.M. Tangen, B.M. McKimmie, Beliefs about error rates and human judgment in forensic science, Forensic Sci. Int. 297 (2019) 138–147.
[23] J. Herskovitz, Influential Texas Panel Recommends Halt to Use of Bite-mark Evidence, Reuters, 2016. https://www.reuters.com/article/texas-bitemark/influential-texas-panel-recommends-halt-to-use-of-bite-mark-evidence-idUSL2N15R00N.
[24] I.A. Pretty, D. Sweet, The scientific basis for human bitemark analyses—a critical review, Science & Justice: Journal of the Forensic Science Society 41 (2) (2001) 85.
[25] Texas Forensic Science Commission 2016–2017, Sixth Annual Report: December 2016-November 2017, 23 Retrieved from:, (2016) . https://www.txcourts.gov/fsc/carousel/texas-forensic-science-commission-sixth-annual-report/.
[26] J.D. Lieberman, C.A. Carrell, T.D. Miethe, D.A. Krauss, Gold versus platinum: do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? Psychol. Public Policy Law 14 (1) (2008) 27–62, doi:http://dx.doi.org/10.1037/1076-8971.14.1.27.
[27] M.H. Ziemke, S. Brodsky, Unloading the hired gun: inoculation efforts in expert witness testimony, Int. J. Law Psychiatry 42–43 (2015) 91–97.
[28] W.C. Thompson, N. Scurich, How cross-examination on subjectivity and bias affects jurors' evaluations of forensic science evidence, Journal of Forensic Science 64 (2019) 1379–1388, doi:http://dx.doi.org/10.1111/1556-4029.14031.
[29] D. McQuiston-Surrett, M.J. Saks, The testimony of forensic identification science: what expert witnesses say and what factfinders hear, Law Hum. Behav. 33 (5) (2009) 436.
[30] J.J. Koehler, If the shoe fits they might acquit: the value of forensic science testimony, J. Empir. Leg. Stud. 8 (2011) 21–48.
[31] J. Kukucka, S.M. Kassin, P.A. Zapf, I.E. Dror, Cognitive bias and blindness: a global survey of forensic science examiners, J. Appl. Res. Mem. Cogn. 6 (4) (2017) 452–459, doi:http://dx.doi.org/10.1016/j.jarmac.2017.09.001.
[32] F. Faul, E. Erdfelder, A.G. Lang, A. Buchner, G* Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences, Behav. Res. Methods 39 (2) (2007) 175–191, doi:http://dx.doi.org/10.3758/BF03193146.
[33] D.M. Oppenheimer, T. Meyvis, N. Davidenko, Instructional manipulation checks: detecting satisficing to increase statistical power, J. Exp. Soc. Psychol. 45 (4) (2009) 867–872, doi:http://dx.doi.org/10.1016/j.jesp.2009.03.009.
[34] B.L. Garrett, G. Mitchell, Forensics and fallibility: comparing the views of lawyers and jurors, West Law Rev. 119 (2016) 621–637.
[35] B.D. McAuliff, M.B. Kovera, G. Nunez, Can jurors recognize missing control groups, confounds, and experimenter bias in psychological science? Law Hum. Behav. 33 (3) (2009) 247–257, doi:http://dx.doi.org/10.1007/s10979-008-9133-0.
[36] J.A. Chorn, M.B. Kovera, Variations in reliability and validity do not influence judge, attorney, and mock juror decisions about psychological expert evidence, Law Hum. Behav. 43 (6) (2019) 542–557, doi:http://dx.doi.org/10.1037/lhb0000345.
[37] J. Kaplan, S. Ling, M. Cuellar, Public beliefs about the accuracy and importance of forensic evidence in the United States, Sci. Justice (2020).
[38] B.H. Bornstein, The ecological validity of jury simulations: Is the jury still out? Law Hum. Behav. 23 (1) (1999) 75–91.
[39] K. Pezdek, E. Avila-Mora, K. Sperry, Does trial presentation medium matter in jury simulation research? Evaluating the effectiveness of eyewitness expert testimony, Appl. Cogn. Psychol. 24 (5) (2010) 673–690.
[40] B.H. Bornstein, J.M. Golding, J. Neuschatz, C. Kimbrough, K. Reed, C. Magyarics, K. Luecht, Mock juror sampling issues in jury simulation research: a meta-analysis, Law Hum. Behav. 41 (1) (2017) 13.
[41] S.M. Kassin, I.E. Dror, J. Kukucka, The forensic confirmation bias: problems, perspectives, and proposed solutions, J. Appl. Res. Mem. Cogn. 2 (1) (2013) 42–52.
[42] J. Koehler, N.J. Schweitzer, M. Saks, D. McQuiston, Science, technology or the expert witness: what influences judgments about forensic science testimony? Psychol Public Policy Law 22 (2016) 401–413.
[43] R.E. Petty, J.T. Cacioppo, Communication and Persuasion: Central and Peripheral Routes to Attitude Change, Springer Science & Business Media, 2012.