J Forensic Sci, May 2009, Vol. 54, No. 3
doi: 10.1111/j.1556-4029.2009.01025.x
Available online at: www.blackwell-synergy.com

Glenn Langenburg,[1] M.Sc.; Christophe Champod,[2] Ph.D.; and Pat Wertheim,[3] B.A.

Testing for Potential Contextual Bias Effects During the Verification Stage of the ACE-V Methodology when Conducting Fingerprint Comparisons*

ABSTRACT: This study was conducted to assess if fingerprint specialists could be influenced by extraneous contextual information during a verification process. Participants were separated into three groups: a control group (no contextual information was given), a low bias group (minimal contextual information was given in the form of a report prompting conclusions), and a high bias group (an internationally recognized fingerprint expert provided conclusions and case information to deceive this group into believing that it was his case and conclusions). A similar experiment was later conducted with laypersons. The results showed that fingerprint experts were influenced by contextual information during fingerprint comparisons, but not towards making errors. Instead, fingerprint experts under the biasing conditions provided significantly fewer definitive and erroneous conclusions than the control group. In contrast, the novice participants were more influenced by the bias conditions and did tend to make incorrect judgments, especially when prompted towards an incorrect response by the bias prompt.

KEYWORDS: forensic science, fingerprints, context bias, cognitive psychology, verification, ACE

The potential (negative) influence of contextual bias is a concern for forensic scientists, especially in the wake of the FBI's high profile erroneous identification to Brandon Mayfield in the Madrid train bombing case (1). Context and confirmation biases were identified as the leading causes of the propagation of the error within the FBI and propagation of the error when the FBI's work was reviewed by an independent expert hired by the defense (2,3).

Judges in the cases of *New Hampshire v. Richard Langill* and *Maryland v. Bryan Rose* actually excluded fingerprint evidence (although *Langill* was later overturned by the Supreme Court of the State of New Hampshire) based in part on the mere potential of error caused by observational biases (4–6). While no error has yet to be shown in either of those cases, the judges were concerned that failure to verify the work of another scientist using blind procedures could produce a similar error as the FBI's erroneous identification in *Mayfield.*

Cognitive biases can come in many forms. At issue in the *Mayfield* case was context bias and confirmation bias. Context bias can be described as a bias due to exposure to extraneous information which is unrelated to the task or decision at hand (7). The following three examples illustrate contextual information that should have no bearing on the task of comparing an unknown fingerprint to a known fingerprint: knowing that a suspect has admitted to being at the scene, knowing that a suspect has a history of similar crimes, or as in the *Mayfield* case, knowing that at least one other qualified expert has already declared the fingerprint evidence a match. Confirmation bias is related to the expectations of the observer (8). The observer tends to see what they want or what they have come to expect, rather than evaluate what is present. An example to illustrate this effect is the assertion that a set of data supports the tester's hypothesis, when in fact the data which do not support the hypothesis have simply been ignored by the tester. The tester only "sees" the data that support his position or belief (9). In relationship to fingerprint evidence, confirmation bias can be seen when an analyst, who is told "this is a match," discounts discrepancies during the comparison in favor of the similarities that support the premise that the images are indeed a match.

Fingerprint specialists apply a process known as ACE-V to compare an unknown fingerprint (often referred to as a "latent print" or a "fingermark") against a known fingerprint exemplar. ACE-V is an acronym representing four stages of the examination process: analysis, comparison, evaluation, and verification. According to the Scientific Working Group on Friction Ridge Analysis, Study, and Technology (SWGFAST), the stages of ACE-V methodology can be described as: analysis—the assessment of the quantity and quality of ridge detail present in an image; comparison—a side-by-side comparison of two images; evaluation—the decision process to declare an individualization, exclusion, or an inconclusive opinion; verification—a subsequent examination of the images by a second examiner resulting in confirmation of the initial examiner's conclusion (10). The present study focuses on potential bias during the verification stage.

Unless a blind testing procedure is invoked during the verification stage, the subsequent examination of the evidence by a second specialist is often performed where the second examiner is aware of the conclusions of the initial examiner. Haber and Haber,

[1] Minnesota Bureau of Criminal Apprehension Forensic Science Services, St Paul, MN.

[2] Ecole des sciences criminelles, Institut de police scientifique, University of Lausanne, Lausanne, Switzerland.

[3] Arizona Department of Public Safety Crime Laboratory, Tucson, AZ.

*Funding for this project (#428-25-03) was made possible by a research grant from the Midwest Forensics Resource Center (MFRC), located in Ames, Iowa.

Received 22 April 2008; and in revised from 25 July 2008; accepted 3 Aug. 2008.

Cole, and Steele have all been critical of this fact (11–13). Likewise, Risinger et al. argued that such extraneous information can easily influence the verifying specialist to conform to the opinion of the initial specialist, even to the extent of overlooking or discounting obvious differences (14). This is precisely what was attributed as the major cause of the propagated error in the *Mayfield* case.

Until *Mayfield*, little attention had been paid to the potential influence of context effects during fingerprint examinations. In 2005 and 2006, Dror et al. reported a series of studies testing context effects during fingerprint examinations (7,15–17). The first study involved novices (college students with no formal training in fingerprint examinations). The study showed that when the comparison trials were difficult and ambiguous (typically when the quality of the mark is limited), the participants were more influenced by increased contextual information. When the comparison trials were relatively straightforward with ample information in terms of friction ridge skin features, it became increasingly difficult to bias the novice participants. The second and third studies involved fingerprint expert participants. The study showed that under varying conditions of bias, an effect could be observed when experts compared fingerprints, particularly when these comparisons were difficult or ambiguous. The fourth paper revisited the previous studies and included a statistical assessment of the data. Further discussion of the Dror et al. studies and the relationship to our results occurs later in this paper.

In light of such data and cases, it is tempting to assume that the appropriate remedy for the fingerprint community is to mandate all verifications be done under a blind testing protocol. Such a procedure would require significant increases in manpower and resources; essentially each case (regardless of the first examiner's conclusion) would be worked twice by independent examiners. A more pragmatic approach was suggested by Champod et al., but without any empirical support for such a policy:

> Systematic blind testing is not necessary during most routine examinations; it is time consuming and unnecessarily consumes significant personnel resources. Rather, a verification structure *should cater to potentially problematic latent prints and cases* (emphasis added) (18).

SWGFAST, in the same vein, also recommends blind testing in limited instances:

> Agencies are encouraged to develop procedures for blind verification. These procedures may be applied to cases where there has been a single conclusion (individualization, exclusion, or inconclusive) to an individual or a complex latent print comparison. Blind verification may also be implemented through random case selection (19).

In the present study, we aim at exploring the issues of contextual bias. We conducted a series of experiments involving a large pool of experts ($n = 43$) and novices ($n = 86$) with a fourfold set of objectives:

1. Determine the effect, if any, the knowledge of the initial examiner's conclusions, identity, or reputation has on the decision-making process of the verifier.
2. Assess variables, such as training, experience, education, demographics, etc. for correlation, if any, between participants and their results.
3. Conduct the experiment on a group of nonprofessionals, as a control and reference for baseline evaluation and comparison against specialists' performance.
4. Based on the results, make recommendations for potential blind testing regimes.

Methods

The experiment with fingerprint specialists (hereafter: "experts") was conducted at the International Association for Identification 91st Educational Conference in Boston, Massachusetts. We chose this venue because it is attended annually by over 1000 forensic identification professionals, many of which specialize in fingerprint identification. A conference room was reserved for the experiment and we solicited for fingerprint experts to participate in an experiment, which was titled "Measuring Variation in Expert Evaluation During Latent Print Comparisons" and a short description of our goals was provided. The description was generic and stated that we were testing "variations in examiner opinions." This was consistent with previous studies conducted by the primary author and was purposefully deceiving. Deception was necessary for testing contextual bias without participant awareness.

The expert pool was separated randomly (through a randomly assigned packet color) into three rooms (and thus three groups: A, B, and C). Groups A, B, and C served as the control group, the low bias group, and the high bias group respectively. At the time of the study, the participants were not told anything about the group in which they were placed, nor did they have any knowledge of what occurred in the other groups.

Each group received a set of six side-by-side comparisons of a fingermark (unknown impression) and a fingerprint exemplar, marked Q1 through Q6. The six comparisons were provided as photographs at 1:1 scale, as 3.5 inch by 5 inch enlargements, and as 8 inch by 10 inch enlargements. A laptop with digital images of each trial and Adobe Photoshop software was provided to each group. Participants wishing to perform comparisons in a digital format were allowed to do so. Participants were provided with a worksheet to provide their evaluation of the comparison. They were instructed to provide an opinion of "individualization" (the images are from the same source to the exclusion of all others), "exclusion" (the images could not have come from the same source finger), and "inconclusive" if neither individualization nor exclusion could be opined. The participants were instructed to provide an explanation if "inconclusive" was chosen. Additionally, the worksheet asked the participant to count the number of minutiae in agreement, the number of minutiae in disagreement, and to rate the quality (clarity) of the fingermark and the quality of the fingerprint.

The images were the same for all three groups. The instructions included definitions of the opinions for conclusions (i.e., individualization, exclusion, and inconclusive) and a consistent counting system for minutiae (e.g., enclosures should be counted as two bifurcations and not as a single minutia). Experiment proctors used a rehearsed script for each group.

The control group received the images with no context information and the experts were asked to provide their opinions and complete the worksheet. The low bias group received the images with a worksheet which provided conclusions for each of the trials. The participants were told that these conclusions were opinions provided by a latent print examiner trained to competency. They were required to state whether or not they agreed with the prompted opinion. The high bias group was provided a similar worksheet as the low bias group; however, this group was told by a prominent, internationally recognized expert in the discipline of friction ridge comparison, that these were his opinions from an actual case. He also provided a copy of an official agency report stating his conclusions and further attempted to persuade his group by providing some analysis commentary for each trial. For example, in Q2, the trial which contained a close nonmatch (look-alike impressions from different sources obtained from a large AFIS database search), the prominent expert instructed "there was an obvious distortion present in this match that any examiner familiar with Ashbaugh's red flag of a 'V-shape' will recognize." Thus, we effectively provided an explanation for the expert participant to discount differences that might be perceived in the comparison.

After all participants had completed the six trials, the participants were reassembled for a debriefing. The true nature of the experiment and the deception were revealed to the participants. What next occurred is critical to interpreting the results that were obtained from this experiment. After revealing the deception, many participants from the bias groups revealed that they had "caught on" and realized our intentions. Several participants stated they were suspicious when the prominent expert began telling them information about the trials. They felt that this was inappropriate for an experiment as it "introduced the potential of some bias." However, once some participants began to perceive differences in a supposed match, they immediately had difficulty accepting the prominent expert's explanations to account for the perceived differences. Several participants also relayed having strong emotional responses at the time and relayed difficulty reaching a conclusion.

Therefore, it is critical when assessing these data to state that we have assumed the following:

- The participants in the bias groups were alert and suspicious.
- The participants in the bias groups were cognizant of a testing environment.
- The participants in the control group had no idea what was occurring in the bias groups.
- No group was aware of the conditions of the other groups.

The experiment was repeated with laypersons with no training or experience comparing friction ridge impressions (hereafter: "novices"). University students at a community college in St. Paul, Minnesota served as the novice groups. These students ranged from 19 to 65 years of age and represented a variety of majors (e.g., criminal justice, computer science, psychology, natural science, and law enforcement). Over the course of three semesters, students attending an elective, introductory forensic science course participated as one of the bias groups (control, low bias, and high bias). The same protocols were followed for each group. Each semester received approximately an hour of friction ridge comparison science and instruction prior to participation in the experiment. Our prominent expert was brought in to deliver the fingerprint lecture for the high bias group. These students were primed by being required to write a paper on the Shirley McKie case and our prominent expert gave a 30-min presentation on his involvement in the case. Incentive was provided in the form of "extra-credit" based on the student's performance.

Our prominent expert was Pat Wertheim, one of the co-authors. Wertheim has been involved in several international, high-profile cases, including the Shirley McKie case from Scotland. It is reasonable to assume that all of the experts that participated in the study were knowledgeable of Wertheim's involvement in the McKie case. To further support that assumption, we scheduled the experiment to follow a presentation of the McKie case at the conference where the expert experimental trials were conducted.

Demographics

Forty-three specialists comprised the expert testing group. There were 20 males and 23 females. There were 15, 12, and 16 experts in the three experimental groups: Group A (control), Group B (low bias), and Group C (high bias), respectively. Experts performed a self-evaluation and classified themselves into one of three categories: "certified," "trained to competency," or "other." There were 21 experts certified by the International Association for Identification (I.A.I.) or another national registry body, 20 experts that declared themselves "trained to competency" and were performing casework, and two participants that comprised a category of "other" (examples include trainee, manager no longer working cases, AFIS-only operator, etc.). The reported years of experience ranged from 1 to 29 years (mean = 11.1 years, SD = 8.0 years) with 41 participants answering this question. Four of the participants received their training outside of the U.S. and were employed as examiners outside of the U.S.

Eighty-six laypersons participated in the novice group. These were separated into the three experimental groups: control (*n* = 31), low bias (*n* = 27), and high bias (*n* = 28). Twenty-nine of the novice participants were male and 57 were female. None had ever analyzed and compared a fingerprint prior to participating in this study.

Results

Figures 1–6 show the images from the six trials that were presented to all participants. Tables 1 and 2 display the conclusions, per trial, for the expert pool and novice pool.

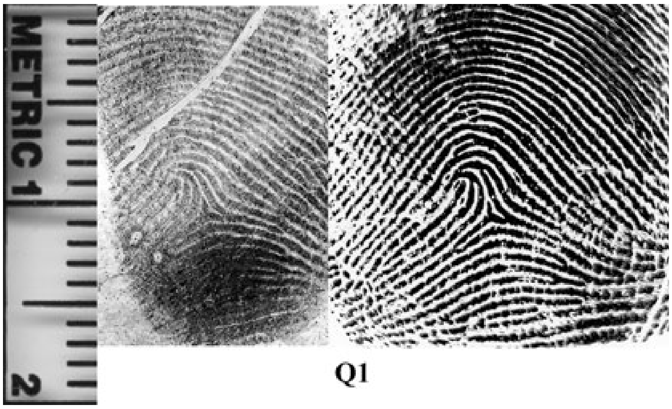

FIG. 1—*Trial Q1 (classified as an easy "same source" trial).*

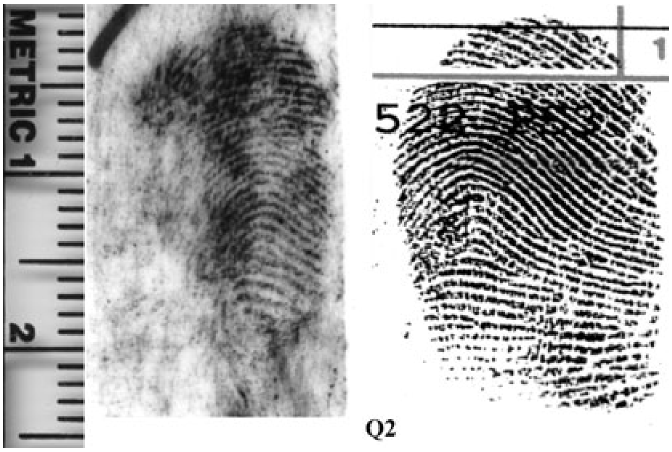

FIG. 2—*Trial Q2 (classified as a difficult "different source" trial). This trial is a close nonmatch resulting from an AFIS database search.*

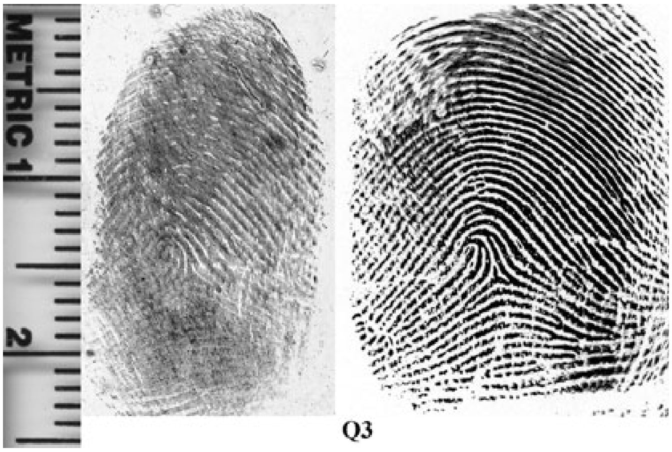

FIG. 3—*Trial Q3 (classifed as an easy "different source" trial).*

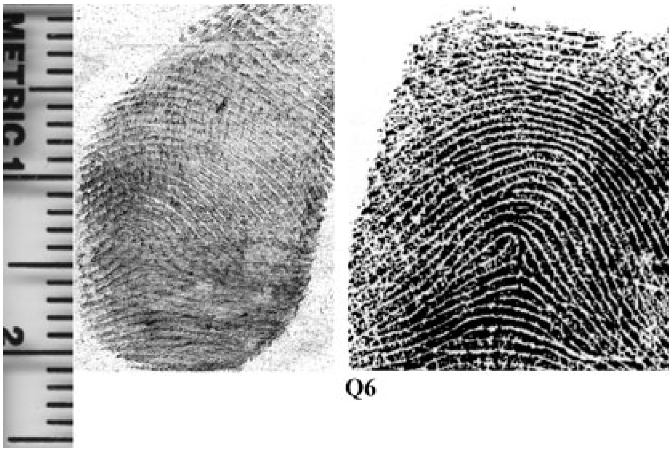

FIG. 6—*Trial Q6 (classified as a medium "different source" trial).*

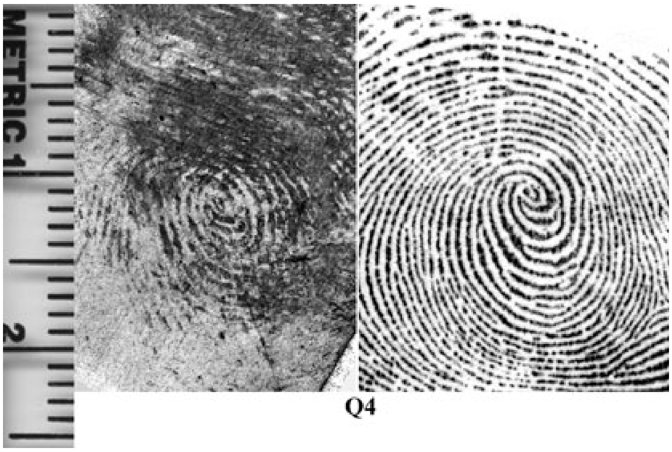

FIG. 4—*Trial Q4 (classified as a difficult "same source" trial).*

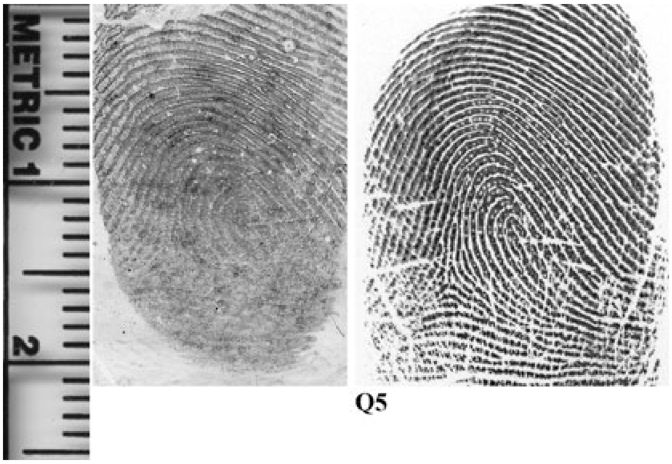

FIG. 5—*Trial Q5 (classified as a medium "same source" trial).*

Discussion

Context Bias Effect

Our primary aim was to test whether there was a measurable effect in the groups that were presented the trials with a context bias. We did observe an effect. Figures 7 and 8 show pooled expert responses. The three trials for which the ground truth was "same source" were pooled together in Fig. 7. Similarly, the three trials for which the ground truth was "different source" were pooled together in Fig. 8. For both sets of data, there was a significant increase in the number of inconclusive responses from experts in the low bias and high bias groups. The percentage of inconclusive responses remained relatively constant between low and high bias groups and relatively constant between responses for "same source" versus "different source" comparisons.

The results of novices were similarly pooled in Figs. 9 and 10. The relative percentage of inconclusive responses remained constant for all three experimental groups for the "same source" trials. It also remained constant for all three experimental groups for the "different source" trials. However, comparing Figs. 9 and 10, there is an increase in inconclusive responses for the "different source" trials. Thus, we conclude that novices had more uncertainty and more difficulty determining an exclusion than determining an individualization. A similar effect has been observed by Kam et al. for novices when performing handwriting comparisons (20–22). In the Kam et al. studies the distribution of errors made by novices was quite disparate for "same source" trials compared to "different source" trials.

A similar disparity between "same source" and "different source" trials was observed for expert participants, but to a lesser degree. The expert participants produced inconclusive opinions for all three "different source" trials, but expert participants produced inconclusive opinions for only one of the three "same source" trials. The difficulty of the three "same source" trials was on a par with the difficulty of the three "different source trials" (i.e., an easy case, a moderate case, and a difficult case).

In Table 3, we can see that the effect of the bias prompt is stronger for novices. In nearly every trial, the relative percentage of expert responses consistent with the bias prompt for the low and high bias groups was either equal to or less than the control group. The exception to this was trial Q6. Here the bias groups had a higher percentage of responses reflecting the bias prompt when compared to the responses of the control. It would appear that the experts were susceptible to the bias prompt (towards inconclusive) in trial Q6.

In contrast, the novices showed a marked increase in relative percentage of responses consistent with the bias prompt for the low and high bias groups when compared to the responses in the control group. The largest jump can be seen in Table 3 for trial Q2

TABLE 1—*Summary of all expert conclusions for the six trials, sorted by experimental group.*

Trial	Ground Truth	Bias Prompt	Group A Control (*n* = 15)	Group B Low Bias (*n* = 12)	Group C High Bias (*n* = 16)
Q1	Same source	Individualization	Individualization = 15	Individualization = 12	Individualization = 16
			Inconclusive = 0	Inconclusive = 0	Inconclusive = 0
			Exclusion = 0	Exclusion = 0	Exclusion = 0
Q2	Different source	Individualization	Individualization = 0	Individualization = 0	Individualization = 0
			Inconclusive = 1	Inconclusive = 2	Inconclusive = 3
			Exclusion = 14	Exclusion = 10	Exclusion = 13
Q3	Different source	Exclusion	Individualization = 1*	Individualization = 0	Individualization = 0
			Inconclusive = 0	Inconclusive = 2	Inconclusive = 2
			Exclusion = 14	Exclusion = 10	Exclusion = 14
Q4	Same source	Individualization	Individualization = 9	Individualization = 3	Individualization = 9
			Inconclusive = 3	Inconclusive = 9	Inconclusive = 7
			Exclusion = 3*	Exclusion = 0	Exclusion = 0
Q5	Same source	Individualization	Individualization = 15	Individualization = 12	Individualization = 16
			Inconclusive = 0	Inconclusive = 0	Inconclusive = 0
			Exclusion = 0	Exclusion = 0	Exclusion = 0
Q6	Different source	Inconclusive	Individualization = 0	Individualization = 0	Individualization = 0
			Inconclusive = 1	Inconclusive = 3	Inconclusive = 2
			Exclusion = 14	Exclusion = 9	Exclusion = 13

One expert participant did not provide an answer for Trial Q6.
*These data denote errors by the participants (opinions contrary to the ground truth).

TABLE 2—*Summary of all novice conclusions for the six trials, sorted by experimental group.*

Trial	Ground Truth	Bias Prompt	Group A Control (*n* = 31)	Group B Low Bias (*n* = 27)	Group C High Bias (*n* = 28)
Q1	Same source	Individualization	Individualization = 27	Individualization = 23	Individualization = 26
			Inconclusive = 0	Inconclusive = 1	Inconclusive = 2
			Exclusion = 4*	Exclusion = 3*	Exclusion = 0
Q2	Different source	Individualization	Individualization = 2*	Individualization = 2*	Individualization = 9*
			Inconclusive = 20	Inconclusive = 21	Inconclusive = 12
			Exclusion = 9	Exclusion = 4	Exclusion = 7
Q3	Different source	Exclusion	Individualization = 5*	Individualization = 5*	Individualization = 0
			Inconclusive = 16	Inconclusive = 11	Inconclusive = 10
			Exclusion = 10	Exclusion = 11	Exclusion = 18
Q4	Same source	Individualization	Individualization = 10	Individualization = 12	Individualization = 17
			Inconclusive = 18	Inconclusive = 14	Inconclusive = 11
			Exclusion = 3*	Exclusion = 1*	Exclusion = 0
Q5	Same source	Individualization	Individualization = 21	Individualization = 17	Individualization = 23
			Inconclusive = 6	Inconclusive = 4	Inconclusive = 4
			Exclusion = 4*	Exclusion = 6*	Exclusion = 1*
Q6	Different source	Inconclusive	Individualization = 0	Individualization = 0	Individualization = 1
			Inconclusive = 16	Inconclusive = 20	Inconclusive = 17
			Exclusion = 15	Exclusion = 7	Exclusion = 10

*These data denote errors by the participants (opinions contrary to the ground truth).

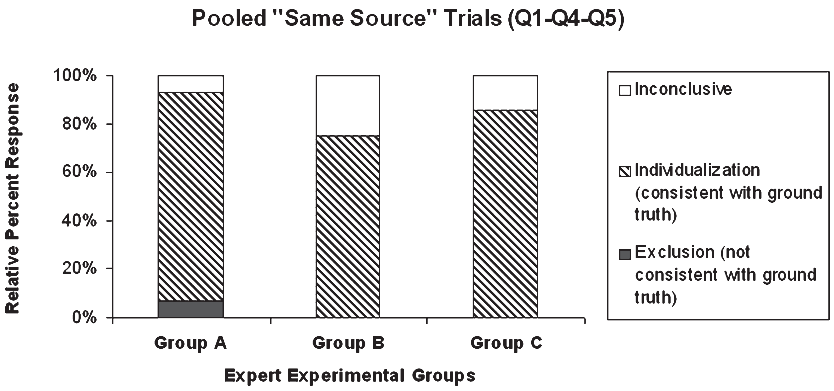

FIG. 7—*Pooled expert trials where the images were from the same source.*

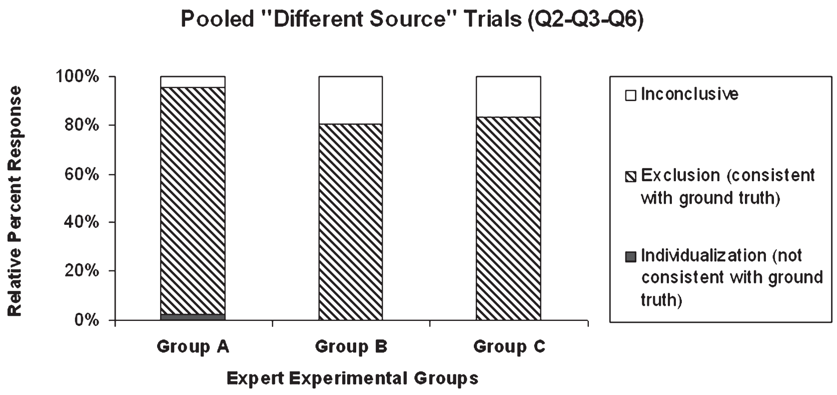

FIG. 8—*Pooled expert trials where the images were from different sources.*

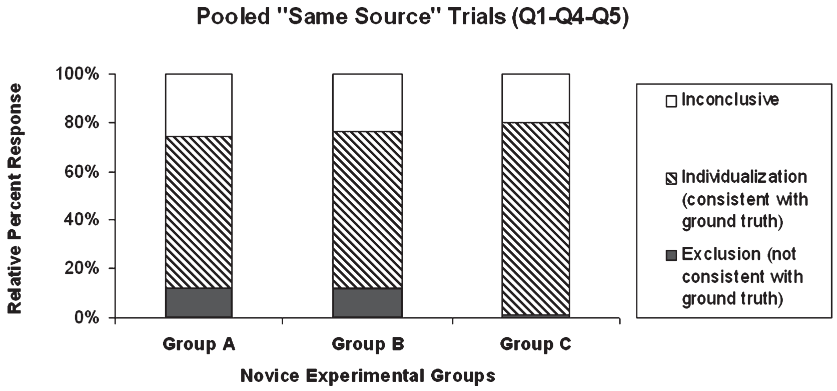

FIG. 9—*Pooled novice trials where the images were from the same source.*

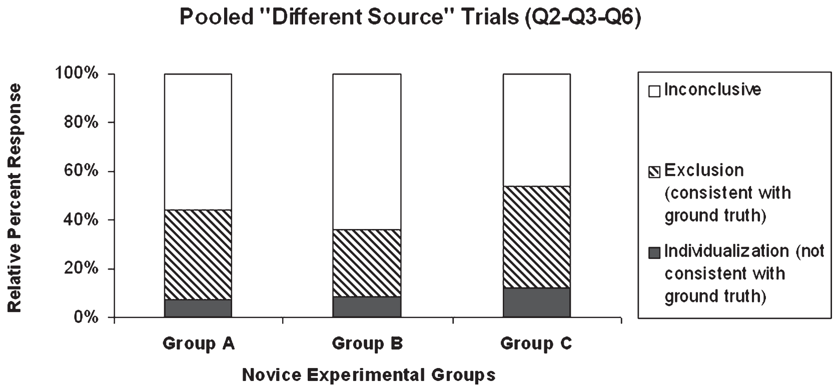

FIG. 10—*Pooled novice trials where the images were from different sources.*

TABLE 3—*Relative percentage of expert and novice responses that were consistent with the bias prompt.*

				Control Group (%)		Low Bias Group (%)		High Bias Group (%)	
Trial	Ground Truth	Bias Prompt	Was Bias Prompt Consistent with Ground Truth?	Expert	Novice	Expert	Novice	Expert	Novice
Q1	Same source	Individualization	Yes	100	87	100	85	100	93
Q2	Different source	Individualization	No	0	6	0	7	0	32
Q3	Different source	Exclusion	Yes	93	32	83	41	81	64
Q4	Same source	Individualization	Yes	60	32	25	44	56	61
Q5	Same source	Individualization	Yes	100	68	100	63	100	82
Q6	Different source	Inconclusive	n/a	7	52	25	74	13	61

(the close nonmatch). Only 6% of the novice responses were consistent with individualization when there was no bias prompt. With a bias prompt from an anonymous report, the number of responses increased a meager 1–7% (we would not call this significant). When the novices were prompted by the prominent expert, the percentage of responses consistent with individualization (an error) increased dramatically to 32%. It is also notable in Table 3 that novices were only minimally influenced, if at all, by the low context bias condition; it was the high context bias condition that had the most significant impact.

We conclude that our prominent expert produced a stronger bias effect in the novice group than in the expert group. We observed a bias effect in the low and high bias expert groups; however, two important distinctions are noted for the expert groups: (i) the bias effect was towards inconclusive responses, rather than definitive exclusion or individualization opinions, as was observed for the novices and (ii) the effect was equivalent in the low and high bias groups—the prominent expert was no more influential on the experts than an anonymous expert's report.

Dror et al. Data

Dror et al. have to date published two experiments involving fingerprint experts and bias (15,16). In the first experiment, the researchers covertly provided identifications to five expert participants that each of the participants had previously made in case work at least 5 years prior to the study. However, when the images were re-presented to the experts, they were presented as the FBI's erroneous individualization to Brandon Mayfield from the Madrid train bombing case. Under this context, three of the experts now provided an opinion of exclusion, one reported an inconclusive opinion, and one reported an individualization (where all five had previously stated they were individualizations).

In the second experiment by Dror et al., the researchers repeated the scheme of the first experiment with six experts under varying bias conditions. In this experiment, there were 48 trials, with eight trials for each of the six experts. Each expert received:

- Two complex cases under no bias conditions.
- Two complex cases under low bias conditions.
- Two simple cases under no bias conditions.
- Two simple cases under low bias conditions.

For these eight trials, four were initially reported by the expert as individualizations and four were initially reported as exclusions. Six of the 48 trials resulted in a response that did not correspond with the initially reported conclusion.

Four of these six incongruous trials resulted in the expert reporting an exclusion, where previously the expert reported an individualization. One of the six incongruous trials resulted in the expert reporting inconclusive, where the expert reported an individualization. The remaining incongruous trial resulted in the expert reporting an individualization, where the expert previously reported an exclusion (this was a complex case re-presented without context bias). Table 4 summarizes all ten instances in the Dror et al. studies where experts changed opinions.

Thus, Dror et al. conducted 53 trials between the two experiments. These trials resulted in 10 trials where a specialist reached a different conclusion than initially reported. Of these 10 instances, it is important to note that nine of the incongruous responses were responses of "inconclusive" or "exclusion," where the initial response was "individualization." Only one trial was from an initial conclusion of exclusion to an individualization, and this trial was not a trial presented under the researchers' context bias conditions.

We argue here that the Dror et al. data are consistent with our data on two grounds. The first is that there is strong evidence that some fingerprint specialists *can* be biased by contextual information. The decision made by a specialist is not necessarily based solely on the ridge detail when comparing images. More importantly the bias effect was most often observed during complex comparison trials.

The second ground is that fingerprint specialists appear to be more susceptible to bias when biased towards inconclusive and exclusion prompts than towards individualization prompts. One reason for this may be in the fingerprint decision-making paradigm and risk mitigation when forming conclusions. Experts are trained to be conservative in their decisions when making individualizations. Penalties for incorrect individualizations can include temporary removal from casework, permanent dismissal from duties, civil law suits, or criminal penalties. The penalties for an erroneous exclusion are not usually as severe. Furthermore, it is difficult in instances of inconclusive opinions to determine if an error did occur. Because of the disparity between the treatment of erroneous individualizations and erroneous exclusions, experts may have developed a susceptibility to bias towards inconclusive and exclusion responses and may be more robust to bias towards individualizations.

Errors

We have chosen to define an error for the present study, as a definitive opinion (exclusion or individualization) that did not reflect the ground truth ("same source" or "different source"). A

TABLE 4—*Summary of Dror et al. experimental results where experts changed initial conclusions to a different conclusion during the experiment.*

Experiment Number*	Expert's Initial Conclusion	Contextual Information Present?	Expert's Second Conclusion	Difficulty of Case
1	Individualization	Yes (high bias)	Exclusion	Difficult
1	Individualization	Yes (high bias)	Exclusion	Difficult
1	Individualization	Yes (high bias)	Exclusion	Difficult
1	Individualization	Yes (high bias)	Inconclusive	Difficult
2	Individualization	None	Exclusion	Difficult
2	Individualization	Yes (low bias)	Exclusion	Difficult
2	Individualization	Yes (low bias)	Exclusion	Difficult
2	Individualization	Yes (low bias)	Exclusion	Difficult
2	Individualization	Yes (low bias)	Exclusion	Not difficult
2	Exclusion	None	Individualization	Difficult

**Experiment numbers 1 and 2 refer to Dror et al. studies:* Contextual information renders experts vulnerable to making erroneous identifications *(11) and* Why experts make errors *(12).*

failure to detect or reach an opinion (inconclusive) has been considered by some as an error (11). In terms of hypothesis testing, a failure to reject or accept the null hypothesis when the null hypothesis does not reflect the ground truth is considered a Type 1 or Type 2 error (respectively). However, there is a problem with the analogy of hypothesis testing and friction ridge comparisons in that a failure to reject the null hypothesis in fingerprint comparisons does not automatically lead to acceptance of the null hypothesis, when "inconclusive" is available as a choice. This creates an interesting gray area. Therefore, without a quantifiable criterion for reaching an opinion of individualization or exclusion, the only acceptable method of evaluating the significance of inconclusive opinions is by expert consensus. Thus, we have chosen not to categorize inconclusive opinions as errors *per se*. Instead, to evaluate the significance of these data, we have compared relative frequencies of inconclusive opinions between bias groups and the control group, between "same source" trials and "different source" trials, and between expert and novice responses.

For the expert groups there were a total of four errors committed. Three of the errors were erroneous exclusions (in trial Q4) and one error was an erroneous individualization (in trial Q3). All four errors were committed by participants in the control group. It should be noted that no participant in the expert groups made an erroneous individualization on the close nonmatch, trial Q2.

It is notable that all four errors were committed by members in the control group. As stated in the Methods section, we have assumed that the control group was not in an "alert and suspicious" state. This has interesting implications. It suggests that one way to reduce errors is to keep experts in an "alert" state. Quality assurance mechanisms, such as random selection of cases for full review, regular (but unannounced) audits, or performance monitoring may achieve this "alert" state. The effect, based on the results of this study, may produce more reliable opinions from specialists. However, it may also be possible that this "alert" state may wear off over time. Analysts may, in effect, develop a tolerance to the "alertness" stimuli. Before implementing such a scheme, further testing is recommended.

The novices committed 24 erroneous individualizations. Seven of these were committed by participants in the control group, seven were by participants in the low bias group, and 10 were in the high bias group. Of the 10 erroneous individualizations in the high bias group, nine were committed in trial Q2, the close nonmatch which was prompted as an individualization. Furthermore, there were 22 erroneous exclusions. Eleven of these were committed by participants in the control group, 10 were by participants in the low bias group, and one by a participant in the high bias group. This is significant given that 'individualization' was prompted by the internationally recognized expert in all three "same source" trials. Thus, only one participant in the high bias group made an erroneous exclusion; there were significantly more errors made by novices in the control and low bias groups.

Training and Experience

Experts were stratified into three categories: (i) trained to competence and performing latent print casework, (ii) certified latent print examiner, and (iii) other (e.g., in training, no longer performing casework due to management duties, AFIS operator only). No statistical significance was observed between trained specialists and certified specialists when the opinions reported or the recorded number of minutiae in agreement were compared. Although there were too few participants in the "other" category to perform a statistical analysis, it should be noted that of the four errors committed by expert participants, two of the four errors (an erroneous exclusion and the erroneous individualization) were committed by participants in the "other" category.

The years of experience reported by expert participants appeared to be significant with respect to the inconclusive opinions reported in the three "different source" trials (Q2, Q3, and Q6) but was not significant in the "same source" trials. In the "same source" trials (Q1, Q4, Q5) the mean experience for the 108 exclusion opinions was 11.9 years (SD = 0.77). For the 14 inconclusive opinions (actually there were 16 inconclusive opinions for these trials; however, two opinions were recorded by participants that did not provide their years of experience) the mean experience was 6.0 years (SD = 1.4). A Mann–Whitney test showed that this disparity in experience was significant ($p = 0.002$). In the "same source" trials, Q1 and Q5 had 100% consensus of opinions reported (individualization) among experts. Only trial Q4 had a variance in opinions reported. Years of experience was not a significant factor with respect to the opinions reported for trial Q4 (Kruskal–Wallis test, $p = 0.955$). There were 40 responses for trial Q4 where the participant reported their years of experience. The mean years of experience reported for each of the three possible responses for Q4 was 9.3 years ($n = 3$, SD = 3.2), 10.8 years ($n = 16$, SD = 8.5), and 11.7 years ($n = 21$, SD = 8.4) for the exclusion, inconclusive, and individualization responses respectively.

Experience has been cited as a cause of expert variation (23–25). However, research conducted by Evett and Williams found no statistically significant correlation between years of experience and willingness to offer an opinion of individualization in complex, same source trials (26). In the present study, our results mimic the findings of Evett and Williams. As demonstrated in Fig. 11, the variable "years of experience" did not appear to be a contributing factor to the opinion provided in the complex, "same source" trial Q4. We suggest several possible explanations for this. The first reason is that "years of experience" does not truly reflect actual years of practice, nor does it accurately reflect the number of cases performed by the specialist. In other words, if the specialist only performs a handful of comparisons per week, they will have accumulated far less actual comparison experience than a specialist performing many comparisons daily over the same number of years. The second possibility is that once a specialist is trained to

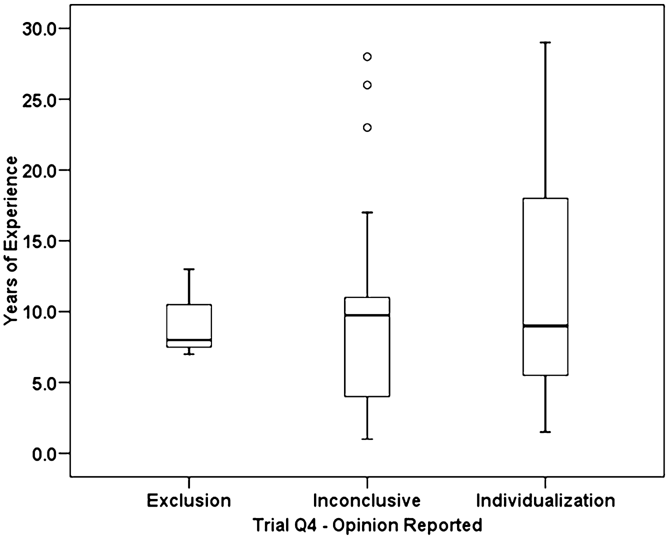

FIG. 11—*Years of experience for fingerprint experts (*n = 40*) versus opinion reported for trial Q4 (difficult "same source" trial).*

competency, the differences (on average) are minimal between a veteran specialist and a significantly less experienced specialist. However, the differences between a specialist trained to competency and a lay person or a trainee are vast. The third possible explanation is that years of experience is a poor predictor of the specialist's conclusions. Over time, one specialist exposed to many close nonmatches may adopt a conservative approach, while another specialist may gain confidence from dealing with many difficult cases over the years. This may cause the latter specialist to "push the envelope" in his or her conclusions. The link between the experience of fingerprint specialists and their conclusions is a fertile area for research because there exists little research that addresses this issue.

Number of Minutiae in Agreement Reported by the Participant

Figure 12 displays the distribution of the number of minutiae in agreement reported by the expert participant for the three "same source" trials, Q1, Q4, and Q5. The data are further stratified by experimental groups in Table 5. While all three groups have a wide range of reported corresponding minutiae, it is notable that in each trial, Q1, Q4, and Q5, the control group has a higher mean and variance for reported corresponding minutiae. A Kruskal–Wallis test shows this effect to be statistically significant at the 0.05 level of significance ($p = 0.021$) when comparing the sum total of minutiae reported for trials Q1, Q4, and Q5. This result suggests that the bias influence extended beyond influencing the opinion (towards "inconclusive") in the bias groups. It suggests that examiners may have also been more conservative when reporting the number of corresponding minutiae. While it is likely that participants in the bias groups were in an "alert" state and therefore gave a more conservative opinion, it is unlikely that they were cognizant of the effect when reporting the number of corresponding minutiae. This effect also has an interesting "chicken-or-the-egg" ramification: did experts "see" less corresponding minutiae because they had already arrived at a conservative opinion ***or*** did experts "see" less corresponding minutiae because of a bias effect and therefore reached a conservative opinion. The research of Schiffer and Champod (27) suggested the former is more likely because the analysis stage was shown to be more robust to context bias effects; however, because the unknown and the exemplar were presented simultaneously to the expert in the present study, we cannot be sure if our findings are consistent with Schiffer or not.

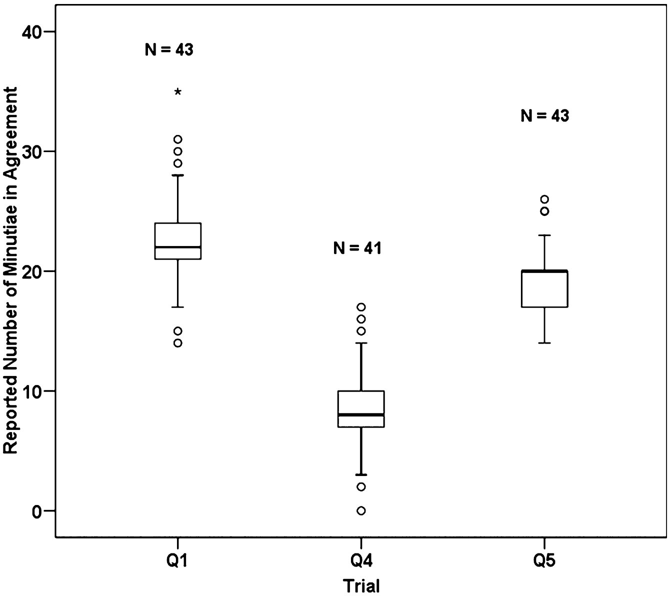

FIG. 12—*Number of minutiae in agreement by the expert group participants (all experimental groups combined) for the three "same source" trials.*

The number of reported corresponding minutiae did not appear significant with respect to the opinion provided by the expert in trials Q1 and Q5, but it was significant in trial Q4. In trials Q1 and Q5, there was 100% consensus among the expert participants; all experts reported individualizations for these trials. Excepting the three erroneous exclusions for the Q4 trial, the expert responses were essentially split (19 inconclusive responses and 21 individualization responses). There was a statistically significant relationship between the number of minutiae reported and the opinion provided by the expert (Kruskal–Wallis test, $p < 0.001$). This suggests that when the expert perceived more corresponding minutiae, they were more likely to report an individualization (see Fig. 13). These data are evidence that in complex cases, such as Q4, the variation in analysis among specialists is critical. In less difficult cases, such as Q1 and Q5, the variation is not critical. Evett and Williams noted similar results in their study (26). We propose that quality assurance mechanisms to reduce variation during the analysis of complex cases be researched.

TABLE 5—*Reported number of minutiae for experts in three "same source" trials.*

	Descriptives							
	n	Mean	SD	SE	Lower Bound	Upper Bound	Minimum	Maximum
Q1—No. minutiae in agreement								
Control	15	24.67	4.546	1.174	22.15	27.18	17	35
Low bias	12	20.50	4.563	1.317	17.60	23.40	14	31
High bias	16	22.38	1.928	0.482	21.35	23.40	20	28
Total	43	22.65	4.058	0.619	21.40	23.90	14	35
Q4—No. minutiae in agreement								
Control	14	9.36	5.048	1.349	6.44	12.27	0	17
Low bias	11	7.64	2.541	0.766	5.93	9.34	3	13
High bias	16	7.81	1.974	0.493	6.76	8.86	4	12
Total	41	8.29	3.459	0.540	7.20	9.38	0	17
Q5—No. minutiae in agreement								
Control	15	20.73	3.058	0.790	19.04	22.43	14	26
Low bias	12	17.75	2.800	0.808	15.97	19.53	14	22
High bias	16	18.06	2.586	0.642	16.69	19.43	14	23
Total	43	18.91	3.061	0.467	17.96	19.85	14	26

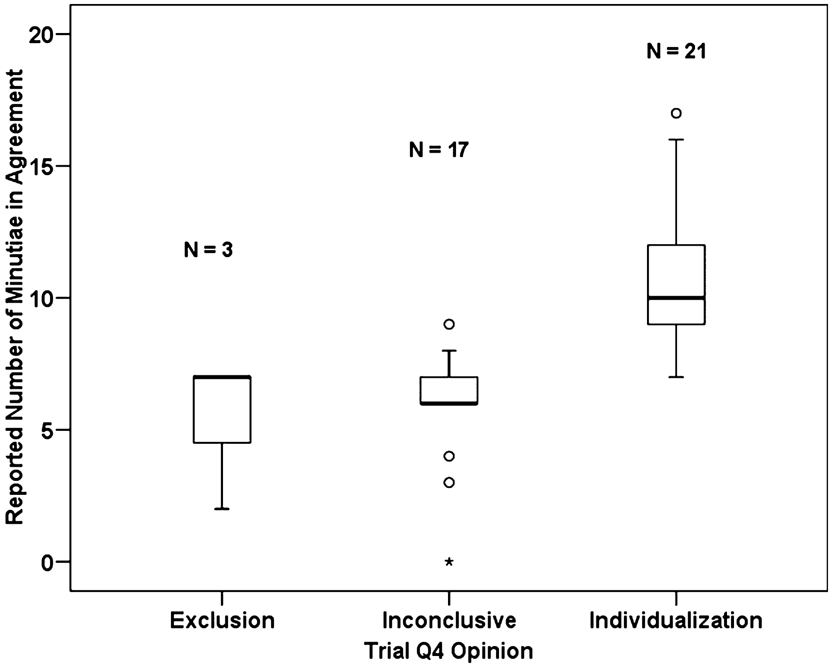

FIG. 13—*Trial Q4 opinion versus the number of corresponding minutiae reported by the expert.*

Other Significant Factors

Each participant completed a survey data sheet to gather background demographic information. The information was generic enough to maintain the anonymity of the participant. We performed General Linear Model multivariate analysis on various factors and standard nonparametric tests for comparing distributions. Factors such as sex, education, country of training or practice, and length of training program did not appear to have a statistically significant effect on the results. Incidentally, 20 of the 43 expert participants possessed a bachelor's degree and 10 possessed graduate degrees. Eighteen of these 30 4-year+ degreed professionals possessed their degree in a physical science (e.g., biology and chemistry). This level of education traditionally was uncommon in the fingerprint profession. The high level of education may have been because of: (i) the location for the experiment (an educational conference), (ii) experiment participation attracts those with a scientific background, or (iii) if this is simply a sign of changing times due to increased scrutiny, hiring practices, and competitive job market in the forensic sciences.

Limitations

The limitations of this study are discussed hereafter, trying to identify clearly what could be robustly inferred from these experimental data.

The Participants were Aware of a Testing Environment

Although the participants did not know they were participating in a context bias study, they were aware that they were being tested. Furthermore, the issue is compounded by the fact that participants in the bias groups became aware and suspicious of the testing environment. It is however noteworthy that the suspicions were only aroused once the fingerprint comparison trials began. Participants were rejecting the notion that Q2 in particular was a match and the bias present was not overriding those bottom-up processes. This in turn led directly to the bias awareness of the participants.

Ultimately, we recognize that the study was not performed covertly, in a case-work like environment. Participants' performance during the study may not have reflected the way they *de facto* conduct comparisons in their case work. This is a strength of the Dror et al. studies. Creating a covert, case-like environment for an experiment is difficult to construct. As a result, Dror et al. had a limited number of participants. Sacrificing the covert conditions, we gained a significant number of participants and were able to perform the same experiment with novices. Comparing the data between novices and experts was very beneficial. Also, the larger sample size allowed us to explore effects such as minutiae reporting, expertise status, sex, and years of experience, etc.

The Experience Level of the Participant was Self-Assessed and Background Varied Significantly

The participants represented various backgrounds and expertise, primarily from the United States. We did not select participants that we could determine were *bona fide* experts. While this type of expert selectivity may be helpful in an experimental design, we felt our sample was more representative of the types of practitioners in the U.S. The practitioners in the U.S. come from a wide variety of training and applied duties.

The Difficulty of the Six Comparison Trials (Q1–Q6) was Calibrated for the Experts, Not the Novices

We selected by consensus of at least three certified latent print examiners, three "same source" trials and three "different source" trials. We selected what we believed to be an easy, medium, and difficult case for each the "same source" and "different source" sets. We used the same images for novices. Therefore, what might have been a medium difficulty trial for an expert, may have been much more difficult for the novice. We believe this was acceptable for purposes of comparing bias effects across experimental groups, such as comparing novice performance in the control group against novice performance in the high bias group. We did not compare experimental group data between experts and novices. Instead, we compared trends within the novice experimental groups against trends in the expert experimental groups. The exception to this was in our error comparison. In this instance, the difficulty of the images will certainly affect the errors committed. Experts made significantly fewer errors than novices. We believe that this demonstrates an important point: the existence of a need for expertise to assess the comparisons.

Participants May Have Come into the Experiment with Preconceived Notions About the Experiment or Experimenter's Expectations

This issue was raised to the primary author during personal communications with Dr. Dror. He suggested that there might have been an issue with the solicitations for expert participation and the faux name of the study: "Measuring Variation in Expert Evaluation During Latent Print Comparisons." Participants may have come with preconceived notions about performance. We do not perceive this to be a problem. In fact, it was orchestrated in this manner with the view that if experts participated in an experiment where the subliminal message was: "We are measuring variation because it is negatively perceived by critics," then there might be some pressure and inherent bias to conform. The effects of conforming would have been apparent in the bias trials and would have resulted in more (not less) errors. Our intent was that this description of the experiment simultaneously created additional, more subtle, context bias and served as a red herring to obscure our deception.

Conclusions

The following conclusions can be drawn from these experimental results:

A contextual bias effect was observed for novice and expert participants when comparing and assessing fingerprints. That bias effect was stronger for novice participants. Experts were more resistant to bias suggestions towards individualization and less so to suggestions towards inconclusive and exclusion. In fact, experts in an alert state (from the bias effect) provided fewer definitive conclusions, whereas experts that were not in the alert state made more definitive conclusions and as a result, more errors. Experts made significantly fewer errors (4 errors) than novices (46 errors).

Experts in the bias groups reported statistically significant fewer corresponding minutiae in "same source" trials. For them, the number of reported corresponding minutiae was a statistically significant factor with respect to the opinion reported in the difficult "same source" trial (Q4). In the easy and medium "same source" trials (Q1 and Q5), there was significant variation in the reported corresponding minutiae, but there was 100% consensus for the experts in the opinions reported.

Finally the variable *years of experience* for the expert was not a statistically significant factor for trial Q4 (the only nonconsensus "same source" trial). Years of experience was a statistically significant factor for the "different source" trials.

The above results, when taken in conjunction with the Dror et al. results, suggest that fingerprint specialists can be influenced by contextual bias information. Therefore, it is important that standard operating procedures and evidence testing schemes reflect appropriate consideration to reduce extraneous context information. However, contrary to some proposals, it may not be required in all cases (i.e., every verification must be performed blindly). Based on our results and the findings of Dror et al., a blind testing regime would best be effective when used in complex cases. Furthermore, we showed that experts appear more susceptible to bias suggestions of "inconclusive" and "exclusion." Blind testing schemes may hence be more wisely employed where an initial expert has reached a conclusion of exclusion. It is recognized although that many laboratories may not have exclusions routinely verified. Those laboratories that do have exclusions routinely verified should consider instituting blind testing when possible to reduce false negatives.

Acknowledgments

We wish to thank the International Association for Identification, and in particular, Jim Gettemy, for giving us valuable conference time and space to conduct these experiments. We would also like to thank Dr. Itiel Dror, David Charlton, Beatrice Schiffer, and several anonymous reviewers for their constructive comments. Finally, and most importantly, we would like to thank the 43 expert participants and 86 novice participants for their time, patience, and assistance in this research.

References

1. Stacey R. Report on the erroneous fingerprint individualization in the Madrid train bombing case. J Forensic Identif 2004;54(6):706–18.
2. Office of the Inspector General (OIG) A review of the FBI's handling of the Brandon Mayfield case. Washington, DC: U.S. Department of Justice, 2006. Available at: http://www.usdoj.gov/oig/special/s0601/PDF_list.htm. Accessed March 4, 2009.
3. McRoberts F, Possley M. Report blasts FBI lab: peer pressure led to false ID of Madrid fingerprint. Chicago Tribune 2004 Nov 14. Available at: http://www.chicagotribune.com/news/specials/chi-0411140299nov14,0,3336405.story. Accessed January 2, 2009.
4. *State of New Hampshire v. Richard Langill*. No. 05-S-1129, January 19, 2007.
5. *State of New Hampshire v. Richard Langill*. No. 2007-300, Supreme Court of New Hampshire, April 4, 2008.
6. *State of Maryland v. Bryan Rose*. Case No. K06-0545, 2007.
7. Dror IE, Peron AE, Hind SL, Charlton D. When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints. Appl Cogn Psychol 2005;19:799–809.
8. Cordaro L, Ison JR. The psychology of the scientist: observer bias in classical conditioning of the planarian. Psychol Rep 1963;13:787–9.
9. Nordby J. Can we believe what we see, if we see what we believe? Expert disagreement. J Forensic Sci 1992;37(4):1115–24.
10. Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST). Friction ridge examination methodology for latent print examiners. ver. 1.01. 2002, Available at: http://swgfast.org.
11. Haber L, Haber RN. Scientific validation of fingerprint evidence under Daubert. Law, Probability and Risk 2007;7(2):87–109.
12. Cole SA. More than zero: accounting for error in latent fingerprint identification. J Crim Law Criminol 2005;95(3):985–1078.
13. Steele LJ. The defense challenge to fingerprints. Crim Law Bull 2004;40(3):213–40.
14. Risinger DM, Saks MJ, Thompson WC, Rosenthal R. The Daubert/Kumho implications of observer effects in forensic science: hidden problems of expectation and suggestion. Calif Law Rev 2002;90(1):1–56.
15. Dror IE, Charlton D, Peron AE. Contextual information renders experts vulnerable to making erroneous identifications. Forensic Sci Int 2006;156:74–8.
16. Dror IE, Charlton D. Why experts make errors. J Forensic Identif 2006;54(4):600–16.
17. Dror IE, Rosenthal R. Meta-analytically quantifying the reliability and biasability of forensic experts. J Forensic Sci 2008;53(4):1–4.
18. Champod C, Margot P, Lennard C, Stoilovic M. Fingerprints and other ridge skin impressions. Boca Raton: CRC Press, 2004;200.
19. Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST). Quality assurance guidelines for latent print examiners. ver. 3.0. 2006, Available at: http://swgfast.org.

20. Kam M, Wetstein J, Conn R. Proficiency of professional document examiners in writer identification. J Forensic Sci 1994;39(1):5–14.
21. Kam M, Fielding G, Conn R. Writer identification by professional document examiners. J Forensic Sci 1997;42(5):778–86.
22. Kam M, Gummadidala K, Fielding G, Conn R. Signature authentication by forensic document examiners. J Forensic Sci 2001;46(4):884–8.
23. Wertheim P. The ability equation. J Forensic Identif 1996;46(2):149–59.
24. Byrd J. Confirmation bias, ethics, and mistakes in forensics. J Forensic Identif 2006;56(4):520.
25. Leo W. Identification standards: the quest for excellence. Calif Identif Dig 1995;12(1):1.
26. Evett I, Williams R. A review of the sixteen points fingerprint standard in England and Wales. In: Almog J, Springer E, editors. Proceedings of the International Symposium on Fingerprint Detection and Identification; 1995 June 26–30; Ne'urim, Israel. Ne'urim (Israel): Hemed Press , 1995;287–304.
27. Schiffer B, Champod C. The potential (negative) influence of observational biases at the analysis stage of fingermark individualisation. Forensic Sci Int 2007;167(2–3):116–20.

Additional information and reprint requests:
Glenn Langenburg, M.S.
Minnesota Bureau of Criminal Apprehension
1430 Maryland Avenue East
St. Paul, MN 55106
E-mail: glenn.langenburg@state.mn.us