Error Rates for Human Latent Fingerprint Examiners

Lyn Haber and Ralph Norman Haber

Human Factors Consultants
and
Department of Psychology
The University of California at Santa Cruz

Abstract

Fingerprint comparison evidence has been used in criminal courts for almost 100 years to identify defendants as perpetrators of crimes. Until very recently, this evidence has been accepted by the courts as infallibly accurate. We review four kinds of available data about the accuracy of fingerprint comparisons made by human latent fingerprint examiners: anecdotal FBI data; published data on the accuracy of consensus fingerprint comparisons made by groups of examiners working in crime laboratories; the proficiency and certification test scores of latent fingerprint examiners tested individually; and the results of controlled experiments on the accuracy of fingerprint comparisons. We conclude that anecdotal data are useless and misleading; consensus judgments of fingerprint comparisons show either indeterminant or quite large error rates; the proficiency and certification procedures in current use lack validity, and cannot serve to specify the accuracy or skill level of individual fingerprint examiners; and there is no published research evidence on error rates. It is impossible to determine from existing data whether true error rates are miniscule or substantial.

Published in Ratha, N.K. (Ed.), (2004). **Advances in Automatic Fingerprint Recognition** (Ch. 17, pp. 337-358)**.** New York: Springer-Verlag.

Address for correspondence: Dr. Lyn Haber, 730 Rimrock Drive, Swall Meadows, CA 93514. Email to haberhfc@telis.org. Telephone: 760-387-2458.

1. Introduction

A fingerprint expert on television who asserts that the prints on the murder weapon match those of the villain is believed, and the bad guy is convicted. This is equally true in a real life courtroom. For almost 100 years, fingerprint evidence has been accepted as fact in court in the United States and other countries. Until very recently, when the fingerprint expert declared an identification, a match between the defendant's prints and trace prints associated with the crime, this assertion went unquestioned by judge, jury and defense attorney.

The fingerprint examiner believes this judgment is fact. Consider the fingerprint expert's testimony in this section of transcript from a trial in 1911:

> Q: In comparing these fingers it is your opinion that the lines in those photographs were made by the same person?
>
> A: I am positive. It is not my opinion. (People of Illinois v. Jennings, 1911).

Today, nearly a hundred years later, practicing fingerprint examiners continue to believe they never could make a mistake, based on their "exact science," and they believe they never do make a mistake. For example: "The fingerprint expert is unique among forensic specialists. Because fingerprint science is objective and exact, conclusions reached by fingerprint experts are absolute and final." (Hazen, & Phillips, 2001).

A growing number of researchers, the press, and legal scholars are questioning the field of fingerprint matching, from its underlying premises to whether there is any scientific rigor whatsoever in the methodology for making the comparisons. For example, "In fingerprint comparison, judgments of correspondence and the assessment of differences are wholly subjective: there are no objective criteria for determining when a difference may be explainable or not." (Stoney, 1997).

The concerns about error rates and scientific rigor have arisen very recently with respect to the fingerprint profession. Court decisions on the scientific requirements for the presentation of expert opinion in court (e.g., Daubert vs. Merrell Dow Pharmaceuticals, 1993; Kumho Tire vs. Carmichael, 1999) have made two novel demands on experts who present testimony. First, experts are required to demonstrate that their opinions are derived from a scientific base, a science which is documented in a research literature and accepted by peers. Second, experts are required to demonstrate knowledge of the error rate associated with the methodology on which their opinions are based. At present, the fingerprint profession insists that fingerprint comparisons are based on an exact science and that competent fingerprint examiners have a zero percent error rate.

In this chapter, we briefly touch on the scientific basis for fingerprint comparisons; then we describe in detail the available data on error rates in making comparisons.

2. Forensic Science of Fingerprint Comparison

Researchers and scientists have raised two general concerns about whether fingerprint comparison is a science. The first stems from the fingerprint profession's focus on the details on actual fingers in ascribing a science to the comparison of prints: the problems presented by latent prints, those found at crime scenes, are blurrily addressed, if at all. The second concern is the profession's failure to develop an explicit forensic science of fingerprint comparisons, specifically, one that defines the transformations that occur in fingerprint patterns when fingers touch surfaces, and defines methodologies for making comparisons.

2.1 Problems with Latent Prints

A typical latent print associated with a crime differs in many ways from the finger itself, and from the fingerprint image taken by trained technicians under controlled conditions in the police station. Each one of these differences serves to diminish, obscure, distort or eliminate information necessary to the comparison process. (1) Size: the latent print is partial, not complete. Typical latent prints contain only about one-fifth of the finger surface contained in an inked or scanned print. (2) Location: some parts of a finger's surface are more informative than others, and the latent print's smaller area may contain little useful detail. (3) The latent print may be deposited on a dirty surface, which obscures critical features. (4) The latent print may be smudged or smeared. (5) The latent print may be overlaid or underlaid by other prints. (6) The latent print may be deposited on a noisy surface, such as wood, which itself consists of ridges and grooves that may be confused with those contributed from a finger. (7) The medium in which the latent print was deposited, such as sweat, water, blood, or oil may interfere with its definition. (8) The amount of pressure and the direction of the pressure between finger and surface produce distortion in the latent print. (9) A curved or irregular surface distorts the patterns of the finger. (10) The procedure used in lifting the latent print usually causes some loss in print detail.

The result of all of these factors is that latent prints almost **always** contain less clarity, less content, and less undistorted information than a fingerprint taken under controlled conditions, and much, much less detail compared to the actual patterns of ridges and grooves of a finger. These transformations between inked or scanned fingerprints and latent prints must be understood, described, and addressed in the methodology for making comparisons (see Ashbaugh, 1999, for further discussion of the problems with latent prints). Further, the impoverished quality of latent prints must be recognized as an inevitable source of error in making comparisons. As one example, the fingerprint examiner profession must explicitly adhere to the practice that more points of agreement be used to make an identification in court when the quality of the latent print is poor, as compared to rich. This follows from common practice throughout all other areas of scientific comparison. When the quality of the stimulus decreases, the likelihood of error increases, so that **more** data are needed to justify a match.

At present, the fingerprint profession neither describes these transformations systematically (but see Ashbaugh, 1999, for an excellent beginning) nor recognizes the loss of information inherent in the latent print as a

critical problem in the assessment of comparison accuracy. The profession persists in the non-responsive claims that fingers are unique and fingerprint examiners do not make errors. This gap between the professional's claims and the problems inherent in the fingerprint comparison task has led researchers, the press, and the legal system to challenge the assertions of an exact science and of a zero error rate.

2.2 The Status of a Forensic Science of Fingerprint Comparisons

Every forensic science requires five components in order to achieve individuation (Haber and Haber, c-in preparation): (1) a description of the patterns of the objects being used to individuate (in this case the patterns of ridges and grooves of fingers); (2) evidence that those descriptions of the patterns can be used to individuate (a demonstration of a uniqueness assumption); (3) descriptions of how the patterns are transformed when deposited in a crime scene, and how each change can be related back to the original pattern; (4) evidence that the descriptions of the trace crime scene patterns (in this case, the latent prints) are also unique to every individual, so that two different people would never leave patterns that could be confused; and (5) descriptions of a tested methodology to carry out comparisons between pattern traces associated with a crime and the object itself (in this case, between latent prints and inked or scanned prints), one that spells out the sequence of steps to be followed, rules to rank the relative importance of different parts of the patterns, the rules to decide when a crime trace does not contain enough reliable information to be used for comparison purposes, and the rules to decide when a difference must be an elimination of an individual rather than a distortion that has arisen from a transformation.

When this forensic science of fingerprint comparison is well described, tested and verified, then the sources of errors in making fingerprint comparisons can be understood, error rates can be determined for different kinds of latent prints and different procedures for making comparisons, and improvements in technology, training, and comparison procedures can be made to reduce the rate of errors. However, at present, a forensic science of fingerprint comparison is neither well described, nor well tested (Haber & Haber, b-in preparation).

For these reasons as well, researchers, the press and the legal system are questioning the magnitude of error rates in fingerprint comparisons.

2.3 Sources of an Error Rate for Forensic Fingerprint Comparisons

When an examiner testifies in court that the latent fingerprint found at the crime scene matches one of the fingerprints of the defendant, the examiner can be correct, or the examiner can be wrong. If he is wrong, the error could have arisen for one of several reasons: (1) Uniqueness: some non-zero probability exists that a second person could have left a latent print that also matches the fingerprint of the defendant, and such an instance occurred in this case. No research has been carried out to demonstrate whether latent prints are unique. (2) Methodology: the procedures followed by the well trained examiner permits an erroneous identification between the crime scene latent print and the defendant's fingerprint, and that occurred in this case. This probability varies as a function of the quality and quantity of detail present. No research has been

carried out to demonstrate whether a latent print from one source finger would be misidentified to an inked print from a different source finger by all competent examiners following a specified method. (3) Human Error: the latent print examiner made a human error through inattention, poor judgment, lack of training, etc. (We ignore for this discussion instances of police falsification of evidence (see Cole, 2001, for examples), or errors in the chain of custody resulting in the examiner comparing a mislabeled latent.).

For legal purposes, only the overall error rate for fingerprint comparisons is of importance. The source does not matter. For a forensic science, it is of critical importance to identify and understand the sources of errors.

The fingerprint profession, represented by the International Association for Identification (IAI) attempts to guard against human error, as the only possible source of error, since the profession denies the possibility of any uniqueness failures, or of a less than perfect methodology. The IAI has suggested standards of admissibility to the profession, standards of training adequacy, standards of years of experience, individual certification of examiners, certification of laboratories in which fingerprint examiners work, and individual proficiency testing of examiners (Scientific Working Group on Friction Ridge Analysis and Study Technology (SWGFAST), 1997). The IAI also recommends that any comparison an examiner judges to be an identification should be verified by a second examiner before the identification is presented in court. Unfortunately, all these standards are advisory, and there is little enforcement in practice.

In this chapter we draw on four areas of evidence and data on human error rates in fingerprint comparisons: anecdotal FBI evidence of a zero error rate; the accuracy of identifications made by examiners working together in crime laboratories; results from proficiency and certification tests of individual examiners; and results of controlled experiments on the accuracy of comparing fingerprints.

3. Anecdotal FBI Evidence of a Courtroom Zero Error Rate

A senior supervisor in the FBI fingerprint examination crime laboratory recently testified under oath in United States Federal Court (Meagher, 2002) that he had never heard of a single instance of an erroneous identification made in court by an FBI examiner. He observed that it would be headline news, and he offered his lack of such knowledge as evidence that during the 35 years of his experience, no FBI examiner had **ever made** an erroneous identification in court.

“Not hearing about an erroneous identification” can be treated as evidence about the frequency of erroneous identifications if, and only if, these errors would be discovered and made available to the public if they occur. As we will show, the entire legal system conjoins to make erroneous identifications undiscoverable. The belief that fingerprint examiners never make erroneous identifications is so strongly held by every party in the criminal justice system that only a concatenation of extremely unlikely circumstances can uncover such a mistake. Defense attorneys, defendants, police investigators, examiners, jurors, and our society at large all believe that fingerprint evidence is infallible.

As a consequence of this belief, each party in the criminal justice system contributes to the difficulty of exposing an erroneous identification.

Historically, fingerprint evidence has been perceived as unassailable by **defense attorneys**, who until very recently did not consider challenging it. This is one reason that there has been little chance of discovering an erroneous identification during trial.

The **defendant** believes that fingerprint evidence will lead to his conviction. An innocent defendant, counseled by his attorney who knows that the prosecution plans to introduce fingerprint evidence, may avail himself of the opportunity to plead guilty to a lesser crime in order to avoid a more severe sentence. Once the guilty plea is recorded, only a fluke can expose the erroneous identification.

Police investigators believe that fingerprint evidence is sufficient to convict the defendant. Once the police have received a report from their crime laboratory that the crime scene prints match a potential suspect, the focus of the investigation switches to collecting further evidence to demonstrate the guilt of the now identified perpetrator. Other potential leads are ignored and other suspects are not pursued. This switch in focus all but removes the chances that an erroneous identification will be uncovered by further investigation.

The **Profession,** represented by the IAI and by the FBI, supports this aura of absolute certainty: “Friction ridge identifications are absolute conclusions. Probable, possible, or likely identifications are outside the acceptable limits of the science of friction ridge identifications” (SWGFAST, 1997). As a result, the profession as a whole has resisted attempts to collect evidence about accuracy. From their point of view, there is no reason to do so: ridge and groove identifications are absolute conclusions and accuracy is always perfect, accuracy is always 100%. The absence of data about accuracy reduces opportunities for both exposure of errors and self-correction.

Jurors believe fingerprint evidence is true. Illsey (1987) shows that jurors place great weight on the testimony of fingerprint experts, and rank fingerprint evidence as the most important scientific reason why they vote for conviction. Mr. Meagher (2002) could not remember an instance in which an FBI examiner made a positive identification and the jury set the identification aside and acquitted the defendant. This widely held belief in the accuracy of fingerprint comparisons does not provide any scientific evidence whatsoever that the fingerprint testimony is correct. The jury’s conviction does preclude discovery of any further evidence of a possible error.

Our **society** makes it difficult for a convicted but innocent prison inmate to conduct an investigation into the “solved” crime, to enlist law enforcement agencies for assistance, or to engage legal help. These obstacles reduce the chances for an erroneous identification to be exposed once it has been used to convict.

For all of these reasons, if an erroneous identification has been made based on fingerprint evidence, it is extremely unlikely that it will ever be discovered. Mr. Meagher correctly could testify that in his memory, not a single instance of an erroneous identification made in court by an FBI examiner has ever been **discovered.** This more reasonable claim provides no evidence that mistakes are never made; only that if mistakes occur, they are never unearthed.

4. Crime Laboratory Consensus Accuracy

Much of the data on fingerprint comparison accuracy comes from joint decisions made by two or more examiners working together in a crime laboratory. Crime laboratories, including the FBI fingerprint crime laboratory, may require that an identification to be presented in court be verified by another examiner of equal or greater experience or rank.

In the accreditation and proficiency testing described below, many laboratories mirror their routine work procedures and have two or more examiners respond jointly to the test comparisons. The response form returned for scoring, like the identification made in court, represents the consensus of several examiners. We present consensus data from four sources.

4.1. Crime Laboratory Accreditation and Proficiency Test Results

In 1977, the American Society of Crime Laboratory Directors (ASCLD) began work on a program to evaluate and improve the quality of crime laboratory operations. An accreditation program was developed, and in 1982 the first laboratories were accredited. As of 1999, 182 laboratories had been accredited, including several abroad (ASCLD, 2001). One requirement of the accreditation process was that examiners working in the laboratory pass an external proficiency test, and, beginning in 1983, ASCLD constructed and administered by mail an annual proficiency test. Each year, each laboratory requesting accreditation was sent a dozen or more latent prints, along with some number of ten-print cards. The latent prints either were selected from actual cases, or were constructed to represent the range of quality found in typical latent prints. The ten-print cards were also selected to be of typical quality. The examiners in the laboratory had to judge whether each latent print was scorable, and if scorable, whether it matched a fingerprint on one of the ten-print cards, or could be eliminated as matching none of them.

Table 1, taken from Peterson and Markham (1995), shows the percent of erroneous responses by year for 1983-1991 on the ASCLD proficiency test used in their accreditation program. The average error rates over these nine years are shown at the bottom of the table.

Table 1: Crime Laboratory Proficiency Test Results for Fingerprint Examiners from 1983-1991 (from Peterson & Markham, 1995). Scores are the percent of erroneous responses. Note: Blank entries occurred in years in which the response was not requested on the test. Percent of respondents who were completely correct was not reported.

Year	Number	Percent Scorable?		% Identification?	
	Labs	Yes	No	Yes	No
1983	24	9%	0%	15%	0%
1984	28	3	7	5	2
1985	37	3	---	7	3
1986	43	1	0	7	2
1987	52	3	---	8	2
1988	62	1	6	2	2
1989	56	1	43	3	0

1990	74	1	1	15	2
1991	88	4	---	---	---
Average Error Scores		**2**	**8**	**8**	**2**

The number of laboratories that returned a response form increased from 24 to 88 over this period. This was a proficiency test of the laboratory as a whole, so the response for each latent presumably was the "laboratory" response. If two latent print examiners worked in the laboratory, the judgment was made by those two examiners working together. If five examiners worked there, all five may have participated together. One or more examiners may have reached a decision and then shown their work to a supervisor, who may have corrected any errors. No information is available as to how many examiners these response forms include, or as to how the examiners reached consensus.

The numbers reported in Table 1 are the percent of erroneous responses for each category of test item (there is no way to determine how many response forms were perfect). The results are fairly consistent over the nine years, so the average scores are representative (bottom row of Table 1). On average, only 2% of the scorable latent prints were erroneously judged to be unscorable, and so were not examined further when they should have been. Only 8% of the unscorable latent prints were erroneously judged to be scorable and were examined further when they should not have been. (No data were reported on the outcome of these erroneously scored prints: whether they resulted in erroneous eliminations or erroneous identifications.) With respect to eliminations and identifications, only 8% of the identifiable latent prints were eliminated (that is, could have been identified but were not), and only 2% of the elimination prints were scored as identifications (that is, were erroneous identifications).

These 2% findings are extremely troublesome. While an individual 2% erroneous identification rate or an individual 2% erroneous unscorable rate may seem negligible, they assume serious proportions in these tests, because the errors result from consensus and not individual independent judgments. To illustrate, assume that two examiners participated in the certification test, performed the comparisons separately, compared their results, and reported an identification only when they both agreed—that is, reached a consensus. Further, assume (as the data show) that 2% of the time they both agreed on an identification when in fact the two prints did not match. Under these conditions, to have a consensus error rate of 2%, each individual examiner acting alone has an average independent error rate of 14% (0.14 x 0.14 = 0.02). If three examiners participate and agree, a 2% consensus error rate implies that each examiner acting alone has an average independent error rate of 27%. The same observations pertain to the 2% scorable latent prints that were not examined when they could have been.

The proficiency tests used by crime laboratories for their ASCLD certification are taken with full knowledge that the fingerprint examiners and the lab as a whole are being tested. Research on double blind testing procedures (see below) show that test score results are inflated when people know they are being tested, and when they and their supervisors know of the importance of the

test results. Test takers are more attentive, follow instructions better, and check and review their work more carefully, all of which improves their scores, compared to their normal level of performance. The 2% consensus error rate for erroneous identifications underestimates the error rates that actually occur in court.

These apparently good results involving consensus in fact reveal substantial individual error rates for examiners working in crime laboratories.

In 1994 the ASCLD Proficiency Advisory Committee contacted the IAI and asked for assistance in the manufacture and review of future testing materials. The IAI contracted with the Collaborative Testing Services (CTS), and, from 1995 to the present, the external latent fingerprint examiner proficiency test used by ASCLD has been administered by CTS, and designed, assembled, reviewed, and authorized by the IAI. Its format still consists of a number of latent prints and ten-print cards. However, all the latent prints are scorable, so the only responses required are identification or elimination. Also, individual examiners who wish to take this test can do so. The summary responses reported by CTS combine consensus reports from laboratories and from individual examiners.

The overall results for the seven years from 1995 to 2001 are listed in Table 2. We constructed the table from the annual summary reports issued by CTS for these seven years (CTS publications 9508, 9608, 9708, 9908, 0008, 0108, 01-516, and 01-517). These are the only reports available, and they do not provide breakdowns by item type. Therefore, it is not possible, for example, to determine the percent of latent prints that should have been eliminated, but were identified (erroneous identifications): only the percent of respondents who made one or more erroneous identifications is reported.

Table 2: Results for seven years of the Collaborative Testing Service Proficiency Examination for Latent Print examiners. Two separate tests were given in 2001. (A year may not add to 100% if a respondent made more than one kind of error.)

Year of Test	Number Taking Test	% All Correct Responses	% One or More Erroneous ID	% One or More Missed ID
1995	156	44%	20%	37%
1996	184	16	3	80
1997	204	61	6	28
1998	219	58	6	35
1999	228	62	5	33
2000	278	91	4	5
2001	296	80	3	18
2001	120	80	2	18

The third column of Table 2 reports the percent of respondents who made no errors. For example, of the 156 respondents to the proficiency test in 1995, 44% made no errors at all. Of the 56% who made errors, 37% failed to make at least one of the identifications. This 37% includes 4% who failed to make even

one identification (CTS 9508). More seriously, 20% made one or more erroneous identifications, averaging 1.6 such errors per respondent. If these comparisons had been presented in court as testimony, one in five latent print examiners would have provided damning evidence against the wrong person. If these are consensus reports, more than half would have testified erroneously.

The 1996 proficiency test results remained poor. Of the 184 respondents, only 16% correctly identified all nine of the identifiable latent prints while also eliminating the two remaining latent prints. The only improvement over the 1995 results was that fewer erroneous identifications were made (3%)—rather, 80% of the erroneous responses were misses -- failures to make correct identifications.

The results for 1997 through 2001 are also reported in Table 2. In general, the number of respondents who achieve perfect performance increases, and the number of erroneous identifications remains small, though 3% to 6% of the respondents continue to make one or more.

These test results, though extremely poor in their own right, are useless to the profession and useless to the criminal justice system, especially the courts, in trying to evaluate erroneous identification rates. There is no control on how the test is administered or timed (it is still distributed and returned by mail); and the only information retained on who takes the test each year is whether the respondent is foreign or from the US, and if from the US, whether associated with an accredited crime laboratory. The anonymous results published each year still do not include any information as to whether the responses came from a consensus or an individual examiner, or the years of experience or amount of training of those who took the test.

The difficulty level of each test, along with its reliability and validity, is not reported. Difficulty level is not even knowable, because a metric of the difficulty of a print has never been developed and tested. Further, the CTS proficiency tests assess only a small portion of an examiner's typical job. If these tests are to measure proficiency in a way that generalizes to performance accuracy in court, they must include an assessment of handling AFIS search outputs, of eliminations as well as identifications, of prints that cannot be scored at all, and of ten-prints that do not match the latent prints.

The CTS, when reporting their test results, now carefully and properly acknowledge that their proficiency tests do not represent the performance accuracy of print examiners in the field: "This report contains the data received from participants in this test. Since it is their option how the samples are to be used (e.g., training exercise, known or blind proficiency testing, research and development of new techniques, etc.), the results compiled in the Summary Report are not intended to be an overview of the quality of work performed in the profession and cannot be interpreted as such…. These Comments are not intended to reflect the general state of the art within the profession." (CTS, 01-516).

4.2 Crime Laboratory Verification Testing

Meagher (2002) provided some information about verification procedures in the FBI fingerprint crime laboratory. However, crime laboratories in general never reveal how many of the identifications made by an examiner working in the

laboratory are given to another examiner to verify. More importantly, crime laboratories, including the FBI's, never reveal how many disagreements occur during the verification process, and what happens when one arises. Hence, while the claim is made by the profession that verification procedures reduce the chance of erroneous identifications, there are no data to support or refute this claim.

Moreover, the verification procedures described by the IAI (SWGFAST, 1997), to which the FBI adheres, are not followed by all crime laboratories in the United States. Some laboratories have no verification procedure for identifications they offer in court.

There are serious problems with these verification procedures in their present form. In response to Meagher's (2002) testimony in United States of America vs. Plaza et al., Arvizu (2002) and Haber (2002) testified that the FBI's procedures were fundamentally flawed as a means to detect errors.

Verifying the accuracy of a result produced by an examiner in a crime laboratory is comparable to auditing the quality control procedures of a water testing laboratory, or to an experimental test of the efficacy of a new drug compared to a placebo. In general, accurate results are obtained if the person being verified, audited or tested does not know that a verification, audit or test is being performed, does not know the specific purposes of the test, and does not know the expected or desired outcome by whoever is administering the procedure. In addition, the persons administering the verification, audit or test should have no stake in its success or outcome, and should not know the correct, expected or required answers. Finally, the verification, audit or test results should be scored and interpreted by an external and neutral body. When any of these procedures is violated, biased outcomes and inflated scores result.

These procedures, widely known as blind testing in the scientific research literature, are currently required for all federal drug testing programs, and for virtually all peer-reviewed, published scientific experiments. To illustrate why verification procedures in crime laboratories are inadequate, Table 3, taken from Haber and Haber (a-in preparation) differentiates three levels of blind testing for verification tasks.

Table 3: Three levels of Blind Testing described by Haber and Haber (a-in preparation).

Zero Blind Procedures

The fingerprint examiner who has made the initial identification:

Knows he is being tested
Knows what he is being tested on
Knows the person who will verify his result
Knows there are penalties for his poor performance

The second print examiner doing the verification:

Knows he is also being tested
Knows the first examiner

Knows what the first examiner concluded and why
Has a stake in the outcome of the test

The Supervisor:
Knows that a verification is required
Knows who made the initial identification
Knows who did the verification
Knows the outcome desired
Has a stake in the outcome

Single Blind

The Fingerprint Examiner who made the initial identificationaaa;
Does not know he is the first one to make this identification
Does not know whether his work will be verified
Does not know who will verify his work

The Second Print Examiner doing the verification:
Knows the same as under Zero Blind

The Supervisor:
Knows the same as under Zero Blind

Double Blind

The fingerprint examiner who made the initial identification:
Knows the same as under Single Blind

The Second Print Examiner doing the Verification:
Does not know another examiner made an identification
Does not have any access to the work of any other examiner on this case
Does not know he is being asked to verify
Has no unusual stake in the outcome of his work

The Supervisor:
Does not know a verification is underway
Does not participate in the analysis of the verification

The research on blind testing demonstrates that double blind test procedures uncover significantly more errors than single blind, and single blind procedures uncover significantly more errors than zero blind (Arvizu, 2002). Unfortunately, crime laboratories that do verify identifications follow zero blind procedures exclusively (Meagher, 2002). Applying those results to verification of fingerprint examiner's identifications, the conclusion reached by the verifier is more likely to agree with the conclusion reached by the initial examiner who first made the identification than if he had done all of the examination himself from the beginning, in ignorance that any one else had worked on the case. The research is clear: under zero-blind conditions, if the first examiner has made an

identification which is erroneous, the second examiner is likely to **ratify** the error, rather than discover it.

Double blind comparisons for verification take more time and increase costs. They are difficult, if not impossible, to implement under the normal operating procedures of crime laboratories today. Crime laboratories need to understand that their present zero blind verification procedures do not eliminate human examiner error from fingerprint comparison testimony, and courts need to understand that zero blind verification does not provide demonstrated amounts of protection for the innocent.

4.3. Results from the FBI External Crime Laboratory Mitchell Comparison Test

In the trial of United States of America v. Byron Mitchell (1999), a latent print examiner testified to an identification between two latent prints lifted from a getaway car and the ten-print card of the defendant. The defendant claimed innocence and challenged the accuracy of the fingerprint evidence. The FBI attempted to demonstrate the scientific certainty of the identification between the defendant's ten-print and the two latent prints found in the car. As part of the demonstration presented at trial, the FBI sent the two latent prints, together with the defendant's ten-print, to 53 different law enforcement agencies around the United States, told them that this request was very important, and asked that their most "highly experienced" examiners determine whether any identifications could be made (see Epstein, 2001, for the detailed instructions and results). This was a unique opportunity for a demonstration of concurrence among experienced examiners.

Thirty nine agencies returned analyses of the prints to the FBI. Nine of them (23%) found that either one or both of the latent prints did **not** match any of the prints from the defendant's ten-print card.

Two issues are embedded in the FBI's failure to find consensus among the highly experienced examiners in the crime laboratories that responded. First, if fingerprint comparison is a "scientific certainty," as the FBI and the IAI claim, then every competent examiner should reach the same conclusion. Here, 30 laboratories found the latent prints were an identification, but 9 found they were not an identification. Where is scientific certainty in the identification of Mr. Mitchell as the person leaving those prints (who, by the way, is still in prison)?

Second, given the nature and the extreme wording of the FBI request, it is most unlikely that the difference of opinion among these highly skilled examiners was due to human error. From our previous discussion, it must therefore be due to either failure of the latent print uniqueness assumption, or that the methodology followed by crime laboratories allows different results for the same comparisons.

The results of the FBI external crime laboratory (Mitchell) comparison test indicate a lack of a forensic science of fingerprint identification.

4.4. Documented Examples of Erroneous identifications

A substantial number of cases in the United States and in England (Starrs, 1998; Cole, 2001) have exposed instances of identification errors made by latent print examiners in court. Epstein (2001) lists more than a dozen. Because

England requires three verifications before an identification can be presented in court, and SWGFAST guidelines suggest at least two in the United States, we include the data on erroneous identifications presented in court as data on the accuracy of consensus judgments.

Discovery of identification error, thus far, has been a matter of chance: someone else confessed, or other irrefutable forensic evidence came to light. Because identifications made by a latent print examiner have not been even questioned until recently, these uncovered errors represent but a few planets in a galaxy.

In each such case in the United States, the IAI has removed certification from the latent print examiner who made the error in court, claiming that any examiner who makes an error isn't competent. Certification has also been removed from the verifier. A "good" examiner never makes a mistake (Cole, 2001). This punitive practice by the IAI ignores fundamental principles of science: no procedure in itself is errorless, and no human being following a procedure is errorless. The IAI, in maintaining that the science of fingerprint comparison is without error, is hindering the profession's capacity to uncover sources of error, to improve performance, and to allow the legal system and society as a whole to make informed decisions about error rates associated with fingerprint evidence.

The documented cases of erroneous identifications in England suggest significant examiner error rates, because, in each one, three verifications of the identification had been made by highly experienced print examiners using the 16-point comparison standard. These cases further demonstrate that a zero blind verification procedure fails to prevent errors even when three skilled examiners check the comparison. Further, given the high level of examiner experience, these errors again implicate methodological problems rather than human error.

The 1999 case against Shirley McKie (Cole, 2001) glaringly illustrated the fallibility of identifications made by expert examiners. Fingerprint examiners from the Scottish Criminal Records Office testified that McKie's ten-print matched a latent recovered at a murder suspect's home; McKie, who was a Detective Constable investigating the murder, claimed she had never entered the home, and that therefore the print could not have been hers. Two American fingerprint experts testified in her defense: experts against experts, in a field where an identification is held to be a certainty!

We have discussed the reasons why documentation of relatively few erroneous identifications made in court cannot stand as evidence that these errors do not occur. We have presented several lines of evidence that fingerprint examiners, working together, do make erroneous identifications. None of the evidence available permits an estimate of the magnitude of the error rates for consensus comparisons. Further, none of these data provides evidence as to the magnitude of these error rates for individual examiners.

5. Individual Proficiency and Certification Tests

We turn now to results from tests administered to individual fingerprint examiners.

5.1. FBI Internal Examiner Proficiency Test Results

Since 1995, the FBI has mandated annual proficiency testing of every latent fingerprint examiner in their employ. The FBI constructs, administers, scores, and interprets their own examinations. The seven years of results (1995-2001) have never been published, but were acknowledged and produced for the first time in a motion the FBI submitted in United States of America v. Plaza et al. (2002). The FBI contended in that hearing (Meagher, 2002) that the results of these yearly proficiency tests show that their examiners do not make errors in court.

Approximately 60 fingerprint examiners took the test each year. The contents of the test varied somewhat from year to year, but always included between 5 and 10 latent fingerprints to be compared to 2 to 4 ten-print cards. The results showed that none of the examiners taking the tests each year over the seven year period made an erroneous identification: the erroneous identification rate was 0%. Three different examiners each missed an identification once in the seven years, a miss rate of less than 1%.

However, contrary to the FBI's contention, these results should not be interpreted as indicating virtually perfect accuracy by FBI examiners when they testify in court. The FBI proficiency test procedures are so fraught with problems that the results are uninterpretable. These problems include difficulty level, unrealistic tasks, and lack of peer review.

The latent prints were of very high quality: clear, distinct, and rich in information content. A leading fingerprint expert and instructor also testified in the same hearing (Bayle, 2002) that the latent prints in the FBI proficiency tests were so easy to identify that his students would pass these tests with only six weeks of training. In the absence of an independent measure of difficulty, a result that nearly everyone gets a high score is vacuous in meaning. There is no way to establish any measure of validity for these results, because the results have no variation. These scores have a zero correlation with supervisory ratings, number of prints examined in the past, years of training, or years of experience.

The tests sampled only a narrow portion of a typical FBI print examiner's workload: none of the latent prints was unscorable; virtually all comparisons were identifications; there were very, very few eliminations; and there were no comparisons made to AFIS search outputs. As the FBI testified (Meagher, 2002) that most of the latent prints sent to the FBI and seen by examiners are unscorable, that most comparisons made in their laboratory were eliminations, and not identifications, and that at least half of all comparisons were produced by AFIS outputs, the proficiency tests neither mirrored the normal workload of FBI examiners, nor sampled the kinds of cases that end up in court.

Finally, great caution should always be exercised in interpreting "in-house" data that are not collected by an independent testing service, and are not submitted to peer review in the publication process. As we mentioned in our discussion of blind testing, the research literature shows that test performance is inflated when examiners know they are being tested, when those administering the tests know the correct answers to the test, and when it is their supervisors who run the testing program and have a stake in its success.

The problems discussed here were reviewed in detail in the defense testimony at the pretrial hearing of the USA v. Plaza (Bayle, 2002; Haber, 2002; Arvizu, 2002). The results of the FBI proficiency tests do not generalize either to an FBI fingerprint examiner's performance on his job, or to the accuracy of the identifications to which attests in court. Their test results, like those of the ASCLD crime laboratory certification test results, are useless to the profession and useless to the courts as an index of erroneous identification error rates.

5.2 Certification Test Results for Individual Examiners from the IAI Latent Print Certification Board

In 1993 the IAI, through its Latent Print Certification Board, began offering certification (and re-certification) to individual latent print examiners. To be eligible to take the certification test, then and now, an examiner must have a minimum of 40 hours of formal training in classification, filing and searching of inked fingerprints, and a minimum of 40 hours of formal training in latent print matters; a minimum of one year's full-time experience in classification, filing, and searching inked fingerprints; and a minimum of two years' full-time experience in the comparison of latent print material. The latent print examiners who take the certification test have both considerable formal training and experience on the job.

Most latent print examiners working in the U.S. today have not taken the certification test (Newman, 2001).

Each year, the certification test contains a number of practical knowledge sections and 15 latent prints that must be compared to a number of ten-prints. To pass the fingerprint comparison part of the test, the test-taker must identify 12 or more of the 15 latent prints correctly, without making a single false identification. In 1993, only 48% of the 762 applicants passed the test. The pass rate has remained at around 50% through 2001 (IAI, 2001). According to the IAI, the section on latent to ten-print comparisons accounted for nearly all of the failures. No data are available as to what percent of the failures resulted from false identifications.

Given the facts that the majority of working fingerprint examiners have never taken the test, that the latent print examiners who do seek certification are already trained and are working fulltime with latent prints, and that about half the applicants for certification fail the test on the basis of poor fingerprint matching skills, the profession simply cannot claim to be performing without error.

A proper proficiency and certification test performs two functions: it permits quantitative assessment of the individual examiner's skill (in this context, accuracy) on the particular tasks examiners perform in their job setting; and it measures examiner skill on the tasks required for their "bottom line," that is, their accuracy when they testify in court. To fulfill the first function, test results must be demonstrably valid: they correlate with years on the job, with supervisory ratings, etc. To fulfill the second, the proficiency test must include the range of tasks examiners typically perform in the course of their work, including elimination prints, unscorable prints, and AFIS outputs. Like the FBI internal proficiency test, the IAI certification test fails to meet either criterion.

5.3 Results from the United Kingdom Review of the 16 Point Comparison Standard

The practice of requiring a threshold number of corresponding matching points for an in-court identification was abandoned officially in the United States in 1973. It prevails in most European countries, and was in use until 2001 in the United Kingdom, where a threshold number of 16 points of comparison was required. In 1988, the Home Office commissioned a review to establish whether the 16-point standard was necessary. Evett and Williams (1996) report the individual examiner accuracy data collected as part of this review.

Photographs of ten pairs of prints, each pair consisting of a latent print and a ten-print card, were sent to fingerprint bureaus in England and Wales. Each bureau was requested to have latent print examiners of ten or more years experience do the comparisons and do them independently. Each of the examiners was asked to make one of four responses for every pair: (1) a full identification, that is, 16 or more points of correspondence; (2) a non-provable (though probable) identification, that is, 8 to 15 points of correspondence; (3) insufficient for an opinion, that is, a latent print that could not be scored; and (4) an elimination, that is, a latent print that did not match the paired ten-print.

The ten pairs were selected from previous Scotland Yard cases. The experts who chose the pairs had decided that all ten latent prints could be scored. They decided that nine of the ten latent prints matched their respective ten-print, of which six pairs were full identifications, sufficient for court testimony, two were non-provable identifications, and one was either a full or a non-provable identification (the experts were split). Only one pair was an elimination.

Of the 130 anonymous responses returned, not one examiner made a single erroneous identification. Of course, there was only one opportunity out of ten to make such an error, so this result by itself does not mean that erroneous identifications are rare among highly experienced examiners. The respondents missed a surprising number of identifications. Half of the respondents judged at least one of the full identification pairs to be non-provable. Even more surprisingly, one of the two non-provable identifications produced only 54% correct (non-provable) responses, with 38% responding that it was unscorable, and 8% declaring an elimination. The authors point out that the results for this latter pair are “puzzling,” given that 8% of the examiners declared an elimination, whereas 54% declared a non-provable (that is, probable) identification. Here was a latent print which some examiners declared was an identification (probable) and some an elimination. Yet, “Examiners consider any identification, provable or probable, as a moral certainty.” (Evett & Williams, 1996, p. 61).

The most startling result found in this study was the tremendous variability in the number of points of correspondence that the examiners reported for each of the nine pairs that matched. Figure 1 shows the results for three pairs for which the correct response was a full identification. If the number of corresponding points between a pair of images of the same finger is based on science, then each of these three line graphs should have all the experts piled up on a common number.

Figure 1

The number of correspondences reported by the 130 expert examiners (data taken from Evett & Williams, 1995), for three of the ten comparison pairs tested.

Consider pair F: only one examiner incorrectly judged it as non-provable, finding only 14 points of comparison; all of the other 129 examiners correctly reported it as a full identification, with 16 or more points of comparison. However, the number of corresponding points for the full identification reported by the examiners ranged from a low of 16 to a high of 56. How can some examiners find only 16 points of correspondence while others find as many as 56, if the selection of points for comparison is scientific? Evett and Williams (1996) conclude that the determination of the individual points for comparison is subjective.

More serious still are the implications to be drawn from pairs B and E in Figure 1. While the correct answer is Identification, 10% of these highly experienced examiners concluded that pair B was not an identification to which they would testify in court. For pair E, the lack of agreement is even more pronounced: half of the examiners failed to score this as an identification. These results are comparable to those of the FBI Mitchell survey, in which highly experienced examiners, who knew they were being tested, came to disparate conclusions. Human error does not seem to be a likely cause of the lack of agreement. The FBI Mitchell survey and the United Kingdom study both suggest that the comparison procedure itself is unreliable, and is applied inconsistently by highly experienced examiners. Because individual experienced examiners reached different conclusions about the same comparisons, the profession cannot claim an “objective and exact science” in which the conclusions presented are “absolute and final.”

The Evett and Williams study is the only published example that assesses the accuracy with which highly skilled examiners made comparisons independently. However, it cannot be used as data about true error rates, nor as a basis from which to generalize to examiner accuracy in court. Four shortcomings are present. The latent prints in this study were judged by experts to be of sufficient quality so they were all usable for comparisons (so examiner accuracy in determining whether a print was scorable was not assessed); nine of the ten comparisons were identifications, with only one an elimination (so little assessment of erroneous identification rates was possible); each latent print was presented already paired with a single known ten-print card (no search was required); and there were no AFIS outputs or other multiply presented candidates to compare. Each of these lacks make the study non-representative of the normal kinds of comparisons carried out by latent print examiners.

For these reasons, the Evett and Williams (1996) report does not provide information about individual examiner error rates of identifications made in court.

6. Experimental Test Results

6.1 Results of Published Experiments on Examiner Accuracy Using AFIS Outputs

There are no published experimental tests on the accuracy with which human examiners compare latent prints to AFIS outputs, nor any on the effect of the number of candidates produced by AFIS on the accuracy of comparison. This is the case even though the FBI estimated (Meagher, 2002) that at least half of all fingerprint comparisons made by examiners involve AFIS outputs.

When a latent print is submitted to a search of a data base by AFIS, the typical output consists of fingerprint images of a number of candidates which the AFIS has found to share the greatest similarity to the latent print. The AFIS output also rates or ranks the similarity of each candidate to the target latent. The examiner, presumably having already analyzed the latent, compares it to each of the AFIS candidates. Examiners typically begin with the highest ranked candidate and continue through the remaining ones, until they find an identification or until they eliminate all the candidates.

Some comparison problems are unique to AFIS output tasks. First, most of the candidates displayed in the AFIS output share a number of similarities by virtue of the AFIS search and comparison algorithms, even though only one, at best, can be a correct identification, and all of the rest must be erroneous. While the human examiner can usually exclude some candidates as obvious eliminations, many require full analysis and comparison to the latent to eliminate them. When an examiner starts with the highest ranked candidate and does not complete a full analysis and comparison procedure for the other possibilities, there is a greater chance that an erroneous identification will be made. Second, some examiners, especially inexperienced ones, ignore proper use of AFIS as a search device only, and rely on the AFIS rankings rather than on their own judgment. This improper procedure can lead to erroneous identifications. Third, to the extent that the candidate prints produced by AFIS are quite similar to one another, the difficulty of the examiner's task is increased. This has two consequences: the examiner may be more likely to rely on the AFIS ranking; and the examiner is more likely to make an error.

For these reasons, experimental tests are needed to demonstrate the accuracy with which examiners make identifications and eliminations using AFIS outputs. Until such data are available, identifications made in court based on AFIS search outputs will continue to pose special concerns about error rates. At present, there are no data.

6.2 Results of Published Experiments on Human Latent Fingerprint Examiner Accuracy and Error Rates

No such experiments have ever been published.

7. Summary

The FBI anecdotal evidence of zero error rates should be rejected in light of all of the factors which make such exposures virtually impossible. The data on examiner consensus accuracy suggest, on the surface, high error rates for individual examiners. However, the proficiency data from ASCLD and CTS are flawed in such fundamental ways that no conclusions whatsoever on examiner accuracy should be drawn from them. The results from the Mitchell case indicate great variability in judgments of comparison. This variability must be addressed: it suggests a lack of rigorous methodology and/or a failure of the uniqueness

assumption with respect to latent prints. Most decisive is the evidence that fingerprint comparison errors have been discovered and erroneous identifications have been attested to in court. The error rate is not zero, but its magnitude is unknown.

The data from the tests of individual examiners, like the consensus data, suggest that examiners may err, but do not permit any estimate whatsoever of the magnitude of the error rate. Nor do the test results generalize to comparisons examiners normally make on the job.

Our careful search of all of the professional research literature turned up not a single experiment on examiner accuracy, either when comparing latent prints to AFIS outputs or when comparing latent prints to ten-prints. Such data simply do not exist, even though examiners have testified in court about their infallible accuracy in making fingerprint comparisons for almost 100 years.

We conclude from this review of the data on fingerprint examiner accuracy that the error rate is greater than zero, and that no estimate of the magnitude of the error rate can be determined. It would be a gross injustice to the profession to draw any conclusions whatsoever on the basis of these data, taken singly or in combination. It is an even graver injustice to convict a defendant on the basis of a comparison technique for which the magnitude of the error rate is unknown. Research is needed to demonstrate the true accuracy levels for the kinds of tasks print examiners have performed when they testify to an identification in court.

.

References

American Society of Crime Laboratory Directors (2001). Report of the Laboratory Accreditation Board. Website www.ascld-lab.org

Arvizu, J. (2002) Testimony on Mr. Plaza's motion to exclude the government's latent fingerprint identification evidence, hearing before Judge Louis Pollak, USA v. Plaza, et al., U.S. District Court for the Eastern District of Pennsylvania, February 24,2002.

Ashbaugh, D.R. (1999). **Quantitative-qualitative friction ridge analysis: An introduction to basic and advanced ridgeology**. Baca Raton, Fl: The CRC Press.

Bayle, A. (2002). Testimony on Mr. Plaza's motion to exclude the government's latent fingerprint identification evidence, hearing before Judge Louis Pollak, USA v. Plaza, et al., U.S. District Court for the Eastern District of Pennsylvania, February 24,2002.

Cole, S.A. (2001). **Suspect identities: A history of fingerprinting and criminal investigations.** Cambridge: Harvard University Press.

Collaborative Testing Services (1995). Certification Examination Testing Program for 1995. Report # 9508; 9608; 9708; 9808; 9908; 0008; 01-517; 01-518.

Daubert v. Merrell Dow Pharmaceuticals (1993). United States Supreme Court, 509, US, 574 (1993).

Epstein, R. (2001). Reply memo of law in support of Mr. Ramsey's motion to exclude the government's latent fingerprint identification evidence. USA v. Ramsey, U.S. District Court for the Eastern District of Pennsylvania

Evett, Z.W., & Williams, R.L. (1996). Review of the sixteen point fingerprint standard in England and Wales. **Journal of Forensic Identification, 46,** 49-73.

Haber, R.N. (2002). Testimony on Mr. Plaza's motion to exclude the government's latent fingerprint identification evidence, hearing before Judge Louis Pollak, USA v. Plaza, et al., U.S. District Court for the Eastern District of Pennsylvania, February 24,2002.

Haber, L., & Haber, R.N. (a-in preparation). The importance of double-blind procedures in forensic identification and verification. To be submitted for publication, winter 2002-03.

Haber, L., & Haber, R.N. (b-in preparation).The accuracy of fingerprint evidence. To be submitted for publication, winter, 2002-03.

Haber, R.N., & Haber, L. (c-in preparation). Forensic scientific theories of identification: the case for a fingerprint identification science. To be submitted for publication, winter, 2002-2003.

Hazen, R.J., & Phillips, C.E. (2001) The expert fingerprint witness. In H.C. Lee & R.E. Gaensslen (Eds.), **Advances in fingerprint technology** (2nd ed.), p. 389-418. Boca Raton: CRC Press.

Illsley, C. (1987). Juries, fingerprints and the expert fingerprint witness. Paper presented at the International Symposium on Latent Prints, FBI Academy, Quantico, VA., July, 1987.

International Association of Identification (2001). Advanced Ridgeology Comparison Techniques Course. Website: www.scafo.org.

Kumho Tire Co., Ltd. V. Carmichael (1999), 236 U.S. 137, 1999.

Meagher, S. (2002). Testimony on Mr. Plaza's motion to exclude the government's latent fingerprint identification evidence, hearing before Judge Louis Pollak, USA v. Plaza, et al., U.S. District Court for the Eastern District of Pennsylvania, February 24,2002.

Newman, A. (2001) Fingerprinting's reliability draws growing court challenges. The New York Times, April 7, 2001.

People of Illinois v. Jennings, 252 Illinois, 534, 96, NE 1077 (Illinois , 1911)

Peterson, J.L., & Markham, P.N. (1995). Crime laboratory proficiency testing results 1978-1991. II: Resolving questions of common origin. **Journal of Forensic Science, 40,** 1009-1029.

Starrs, J. (1998). More saltimbancos on the loose?: Fingerprint experts caught in a whorl of error. **Scientific Sleuthing Newsletter** (Winter).

Stoney, David A.(1997). Fingerprint Identification. In: David L. Faigman, David H. Kaye, Michael J. Saks, & Joseph Sanders: Eds. **Modern Scientific Evidence: The Law and Science of Expert Testimony**, St. Paul, Minn: West.

Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) (1997). Proposed SWGFAST guidelines. **Journal of Forensic Identification, 47,** 425-437.

United States of America v. Plaza et al. (2002) U S District Court of the Eastern district of Pennsylvania.

United States of America v. Byron Mitchell (1999). U S District Court of the Eastern District of Pennsylvania.

Acknowledgement: The world of highly trained, very experienced fingerprint examiners is a small one and the examiners work under great pressure and overwhelming demands. Several, nevertheless, gave us many hours of their time. We gratefully acknowledge the help and education they provided us. We are profoundly impressed by their professionalism, level of expertise, and commitment to accurate performance.

Biography: Lyn and Ralph Haber are both partners in Human Factors Consultants (www.humanfactorsconsultants.com), a private consulting company. They also hold adjunct appointments at the University of California at Santa Cruz (Psychology). Lyn is trained as an experimental linguist (Ph.D., University of California at Berkeley, 1970) and Ralph as an experimental psychologist (Ph.D., Stanford University, 1957). They have published research and provided testimony on human factors issues in the legal system since 1980. They have recently published a number of papers on the accuracy of eyewitness identifications.