Scientific Validation of Fingerprint Evidence Under Daubert

Lyn Haber and Ralph Norman Haber
Human Factors Consultants
Address correspondence to Lyn Haber, Ph.D., Human Factors Consultants
313 Ridge View Drive, Swall Meadows, CA 93514
Telephone 760-387-2458; Fax 760-387-2459
Email lhaber@humanfactorsconsultants.com

Abstract

When a scientific method is used by an expert to reach a conclusion offered in court, the Frye ruling in 1923 and particularly the Daubert ruling in 1993 requires that the method itself has been shown to be valid. When applied to fingerprint methods, valid means accurately distinguishing between pairs of prints made by one and by two donors. Courts have ruled uniformly in more than 40 Daubert hearings since 1999 that fingerprint evidence rests on a valid method, referred to as the Analysis-Comparison-Evaluation-Verification (ACE-V) method. In this article, we discuss the scientific evidence needed to document the validity of ACE-V. We describe examples of experiments that would provide this evidence, and review the available published research. We briefly describe the testimony presented by fingerprint examiners in these hearings, intended to show that ACE-V meets the Daubert criteria for validity. We analyze evidence for the validity of the standards underlying the conclusions made by fingerprint examiners. We conclude that the kinds of experiments that would establish the validity of ACE-V and the standards on which conclusions are based have not been performed. These experiments require a number of prerequisites, which also have yet to be met, so that the ACE-V method currently is both untested and untestable.

Introduction

A fingerprint comparison was first used as evidence against a defendant in 1892 in Argentina. The first case in the United States was 1904 (Cole, 2001). There are no records of how many criminal defendants have been charged and how many convicted in federal and state courts in the United States since 1904 based totally or partially on fingerprint evidence. Speculative estimates point to over 100,000 indictments in the United States in 100 years (Cole, 2005, suggests even more). Most fingerprint cases escape the scrutiny of a trial. The defendant pleads guilty (sometimes to a lesser charge that may not involve the fingerprint evidence) and waives trial. Altschuler (2005) estimated that 95% of such felony cases are adjudicated in this fashion. Fingerprint evidence has been accepted virtually without challenge or question over this 100 year history.

The United States Supreme Court's Daubert ruling (Daubert v. Merrill Dow Pharmaceuticals, 1993) resulted in a barrage of challenges to the admissibility of fingerprint evidence in federal courts (and in state courts using comparable criteria). The FBI's Onin website (Legal Challenges to Fingerprints, 2005) lists over 40 Daubert and state court rulings in criminal cases since the first one (US v. Byron Mitchell, 2000). The basis of these challenges by the defense has been that the primary method used to compare fingerprints—the Analysis-Comparison-Evaluation-Verification (ACE-V) method--lacks the scientific documentation of its validity required under Daubert. Every court has ruled to admit fingerprint evidence, based on fingerprint's 100 year history of accepted practice, and on the assertions of fingerprint examiners that there is adequate scientific support of the method.

A number of legal scholars have recently published analyses of these Daubert rulings (e.g., Cole, 2004; Epstein, 2002; Faigman, Kaye, Saks, & Sanders, 2002; Saks, 2003). They focus primarily on what they consider to be misinterpretations of the legal meanings of the Daubert criteria, made both by the government and by the Daubert courts themselves. In the present article we focus explicitly on the appropriate scientific basis to demonstrate validity. We write this article as research scientists, and describe the available evidence, the kinds of evidence needed, and experiments that would provide that evidence. We conclude that the ACE-V method has not been tested for validity, and until the necessary work is performed to quantify the method and insure that examiners are using the method correctly and consistently, the method cannot be validated.

Validity and Reliability

Daubert courts often refer to the reliability of scientifically based evidence; we discuss evidence of the validity of the ACE-V method. The distinction is important because a reliable method can consistently produce the same wrong answer, whereas a valid method consistently produces the true answer.

A method is valid when its application produces conclusions in agreement with ground truth. In the case of fingerprint comparisons, ground truth is either certain knowledge that the latent and suspect's fingerprints were made by the same person, or certain knowledge that they were made by two different people. The amount of validity of a method is usually expressed as the converse of an

error rate: the percent of time that the conclusion, based upon rigorous application of the method, agrees with ground truth. If scientific evidence shows that the application of the ACE-V method to produce fingerprint individuation evidence results in conclusions that agree with ground truth with a high probability, the method has met the good science requirements of the Daubert decision by the US Supreme Court.

Evidence can also be assessed for its reliability (Cole, 2006). A reliable method is one that produces, for the same comparison, the same result every time it is used, both by many examiners comparing the same set of latent to suspect prints, and by the same examiners (unknowingly) repeating a comparison they had made previously. The amount of reliability is usually expressed as a correlation: the amount of agreement between repeated uses of the method under comparable conditions. A high correlation of reliability indicates that experts using the method reach the same conclusion nearly every time.

Daubert should not be directly concerned with reliability. A highly reliable method (producing the same result every time) may still be wrong, that is, be invalid and disagree with ground truth. The flat earth belief was held reliably by nearly every observer for centuries and they were all wrong. However, a method cannot be valid if it is unreliable. If the method produces varying results each time it is used, some of those results are incorrect, and hence the method is invalid. Therefore, the scientific focus to meet the Daubert good science requirements should be on validity—the agreement of the method's conclusions with ground truth.

The Steps of Our Analysis

We divide our analysis of evidence for the validity of the ACE-V method into six parts.

First, we describe the four stages of the ACE-V method.

Second, we turn to the kinds of evidence needed to demonstrate the validity of the ACE-V method and describe a prototypic experiment that would provide such evidence, similar to experiments assessing the validity of any method. As part of that description, we show that a search of the research literature fails to uncover any instance of such an experiment applied to ACE-V, and we show that the experimental design requires a set of prerequisites which have yet to be met.

Third, we describe and respond to the arguments presented by the government that ACE-V has already been tested and that it exhibits a zero error rate.

Fourth, we analyze the standards that underlie each of the four conclusions examiners draw, based on application of the ACE-V method. We describe the kind of evidence needed to demonstrate the validity of each standard, and for each, describe an experiment that could provide such evidence. From the published literature, we document evidence that application of the standards as presently practiced produces highly variable and therefore invalid results.

Fifth, we discuss the scientific implications of the courts' reliance upon fingerprint examiners rather than research scientists for evaluation of the validity of the method.

Sixth, we conclude that there is no scientific evidence for the validity of the ACE-V method, and, that until the prerequisites for specifying the method and its application are met, the ACE-V cannot be tested for its validity.

I. The Analysis-Comparison-Evaluation-Verification Method

Fingerprint examiners argue that there are unique and permanent combinations of features on the skin of fingers (and palms and feet, though we confine ourselves in this article to fingers). Further, they argue that images of these patterns (called fingerprints) can be used to individuate people by the proper application of a comparison method. When a perpetrator of a crime touches a surface with a finger and leaves an image of the unique pattern from that finger (called a latent fingerprint), that latent fingerprint image can be found, lifted, and compared to the images of the fingers of a suspect (called exemplar fingerprints, and usually recorded by a trained technician on a ten-print card). Following a method of fingerprint comparison such as the ACE-V, and depending on the examiner's training and experience in the comparison method, the examiner can offer an opinion about ground truth: whether the crime scene latent fingerprint was made by the suspect or by someone else.

The FBI claims in Daubert hearings (Meagher, 1999) that all examiners now use the ACE-V method to make these conclusions, and that there are no other methods in use by fingerprint examiners today. Ashbaugh (2005b) makes a similar claim. Although alternative methods are mentioned in textbooks (Olsen & Lee, 2001), practicing examiners uniformly refer to the fingerprint comparison method they employ as the ACE-V. Therefore, we restrict our discussion to evidence for the validity of the ACE-V method.

A Description of ACE-V

Because the ACE-V method may not be familiar to some readers outside the fingerprint profession, we provide a brief overview. However, neither the International Association for Identification (IAI) as the professional organization of fingerprint examiners, the FBI, nor any other professional fingerprint organization has provided an official description of the ACE-V method, so our description is based on the most detailed of the published sources.

Huber (1959, 1972) first described the structure of this method, which he applied to every forensic identification discipline, but without suggesting a name. The classic FBI Science of Fingerprints (1958, 1988) contains only a few pages on comparison procedures, but neither refers to that method as ACE-V nor distinguishes among its different steps. Ashbaugh (1999) provides much more detail and examples, in what has become the most influential textbook available. Champod, et al. (2004) offers an even more precise description, spelled out in the form of steps in a flow chart. P. Wertheim (2002) also has a relatively detailed description, which he has used as a model in his training courses. Beeton (2001) gives a shorter version, contrasting some of the differences between Ashbaugh (1999) and P. Wertheim (2002); and Triplett and Cooney (2006) also comment on some of the differences among the accounts.

What follows is a theoretical description, distilled primarily from the authors cited above, and from our own training in IAI-sponsored latent fingerprint courses. We know from examiner testimony offered in Daubert hearings and in trials involving testimony from fingerprint examiners that most practicing fingerprint examiners deviate from this description. (We consider below the implications of the facts that there is no agreed-upon description of the ACE-V method in the fingerprint profession, no professional body has approved any one description as the official ACE-V method, and that individual examiners vary in their practice.)

Analysis stage. The Analysis stage begins when a fingerprint examiner looks at a latent fingerprint and decides whether it contains sufficient quantity and quality of detail so that it exceeds the standard for value. If the quantity and quality of detail does exceed the value standard, then the examiner continues the analysis. If the value decision is negative, the latent fingerprint is not used further. The majority of latent prints found at crime scenes are rejected as of no value (Meagher, 2002).

If the fingerprint examiner continues Analysis of the latent print (he has not yet seen the suspect's exemplar prints), he uses the physical evidence contained in the latent print and that produced by the crime scene investigation to determine which finger made the print, the nature of the surface on which it was deposited, the amount and direction of pressure used in the touch, and the matrix (such as sweat) in which the ridge details of the finger were transferred onto the surface. This analysis is necessary to specify each of the sources of distortion in the latent print that causes the inevitable differences between the latent fingerprint and the patterns of features found on the skin (and on the exemplar image).

The examiner then chooses one feature-rich area of the latent print (preferably near a core or delta). Within this area, he selects the particular features along the various ridge paths in the latent print, in their spatial locations relative to one another, to use to start the comparison between the crime scene latent and the suspect's exemplar prints.

Comparison stage. In the Comparison stage, for the first time, the examiner looks at the suspect's ten exemplar fingerprints. He starts with the most likely finger of the exemplar image, based on what he found during the analysis of the latent print. The examiner goes to the same area of the suspect fingerprint that he had selected in the latent print to determine whether the same patterning of features occurs. If it does not, the examiner concludes that that finger of the suspect cannot be the finger which made the latent print: an exclusion of that finger. He then goes to the next finger of the exemplar image, and repeats this process. If all ten fingers can be excluded as the source of the crime scene latent print, the examiner excludes the suspect as the donor.

If the same pattern of features initially noted in the latent print is found in the corresponding area of one of the suspect's exemplar prints, the examiner goes back to the latent print, selects another area and locates the features there and their relative positions to each other. Then the exemplar is again examined in the new area, to determine whether the corresponding features are also

present. This latent-to-exemplar comparison (always in that order) continues until all of the features in the latent print have been compared for agreement of features in corresponding locations in the suspect's exemplar. If substantial agreement is found, the examiner goes to the Evaluation stage.

Throughout Comparison the examiner keeps track of every failure to find a correspondence between the latent and the suspect fingerprint. Any failure in agreement that cannot be accounted for by one of the distortions previously described and labeled in the Analysis stage is a necessary and sufficient condition to exclude the suspect as the perpetrator with the application of the one-unexplained-discrepancy standard. The most common conclusion of the Comparison stage is exclusion (Meagher, 2002).

Evaluation stage. In Evaluation, the examiner applies a sufficiency standard to the amount of corresponding agreement between the latent and the exemplar that dictates his conclusion. If the amount of corresponding agreement exceeds the sufficiency standard, then the examiner concludes that the crime scene latent print can be individuated to the suspect. If the amount of agreement does not exceed the standard, then the conclusion is neither an individuation nor an exclusion—an inconclusive conclusion. Two kinds of sufficiency standards obtain. The first is numeric, in which the amount of agreement is stated as a number, and the threshold for sufficiency is determined by the profession or the crime laboratory. The second is experiential, based on the individual examiner's training and experience.

Verification stage. Verification is employed in larger laboratories for cases in which an examiner has concluded individuation. A second examiner confirms the conclusion of the first. A verification standard describes the rules by which a verifier is selected, informed of the history of the latent print's comparisons, reports his findings, and how conflicting conclusions are resolved.

Scoring the Accuracy of the Four Conclusions

In the overview of ACE-V presented above, the examiner has made four kinds of conclusions: value, exclusion, individuation, or inconclusive. These are described in two reports by the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST, 2002a, b). The classification of these four conclusions as either correct or incorrect requires knowledge of the ground truth. The accuracy of the method in reaching true conclusions, or its converse, its error rate, is critical to admissibility under Daubert. We review here the meaning of correct or incorrect for each of the four conclusions. We then present the evidence for ACE-V accuracy with respect to these four conclusions.

Ground truth. Ground truth is certain knowledge that the latent and an exemplar fingerprint came either from the same donor or from two different donors. Ground truth cannot be known in case work, and therefore, research using results from case work cannot be used to establish the validation of the method. In case work, the purpose of investigation, fingerprint comparison, indictment and trial is to find out as much as possible about ground truth. Neither the examiner's opinion nor the jury's verdict is ground truth, though both (or neither) may be consistent with it.

Classification of the four ACE-V conclusions. Table 1 shows the classification of each of the four conclusions as a function of a known ground truth.

Insert Table 1 near here

Consider the first two lines of Table 1. The no value and the inconclusive conclusions always miss the correct answer. Whether the prints came from one donor or from two, neither of these conclusions agrees with ground truth. When an examiner reaches either of these conclusions, either an innocent suspect still remains at risk of indictment and conviction, or a guilty perpetrator still remains at large.

Consider the last two lines of Table 1. The exclusion conclusion is correct when the two prints are from two different donors (and an innocent suspect is released from suspicion), and erroneous when the same donor made both prints (and a guilty perpetrator remains at large). The individuation conclusion is correct when one donor made both fingerprints (and indictment and conviction of the perpetrator is likely), and erroneous when two different donors made the two fingerprints (and an innocent suspect is at risk of indictment and conviction).

We now turn to evidence of the validity of the ACE-V method to reach conclusions that agree with ground truth.

II. Is the ACE-V Method Valid?

Demonstration that an expert used a scientifically validated method (one that has been shown to produce conclusions that agree with ground truth) is intended to assure the court of the accuracy of that method used in the instant case. Similarly, the published error rate information informs the court of the amount of confidence that can be placed in a conclusion based on the method used to reach that conclusion. The validity and the error rate of the method concern the error inherent in the method, as distinct from evidence of any particular individual practitioner's accuracy in applying that method. We return in a later section to arguments that proficiency testing of examiners provides evidence of validity of the method itself.

In several Daubert hearings concerning forensic methods other than fingerprints, such as polygraphs (US v. Scheffer, 1999) or ear prints (Washington v. Kunze, 1999), the court has rejected a method specifically on the grounds that the validity of the method has never been demonstrated scientifically.

An Experimental Design to Test the Validity of the ACE-V Method

A test of the validity of a methodology (such as the ACE-V) requires an experiment. Because the validity of the *method* is being tested (the probability that conclusions based on the method agree with ground truth), all other potential sources of error must be controlled or eliminated from that experiment.

The subjects are skilled examiners who each compare a number of pairs of prints for which the ground truth is known. For each pair, the examiners apply the ACE-V method, they document the application of the method at each step by completing a response form and report, and they state the conclusion supported by the method. Ideal working conditions are present during the experiment.

The validity of the method (the converse of its error rate) can be computed from the amount of agreement found between the conclusions produced by the

method and ground truth, as in Table 1. All deviations from perfect validity in these results (an error rate greater than zero) can be directly attributable to the method, and not to other possible sources of error.

Prerequisite Conditions

With respect to ACE-V, the following four prerequisite conditions must be met in order to perform the validity experiment (see Table 2).

Insert Table 2 here

Description of the ACE-V method. Definitions of each step in the ACE-V method must be provided. To do this, the profession has to write and then adopt a manual describing in detail the steps of the ACE-V method.

The ACE-V method has yet to be officially described and endorsed as an agreed upon method. The published versions of it differ significantly. Each of the many calls within the profession for the production of a complete, official description of the method implies its absence. For example, the Interpol created a working group in 2000 to define a method for fingerprint identification, and issued a report urging its member countries to adopt a common method (Interpol, 2005). A recent FBI report included a demand for a complete description of the method (Smrz, et al., 2006).

At present, the term “ACE-V” refers to some methodological steps which have not been well described, and to differing procedures. Until the method is specified and endorsed, there is no method to test.

Report form for the ACE-V method. Documentation that the examiners are using the ACE-V method in the experiment must be provided. To do this, the profession has to write and then adopt a report form that examiners complete that shows that each step is followed.

At present, examiners using the ACE-V method in this country are not required by anyone to record the steps they took to reach their conclusion. Their reports provide only their conclusions, based on the prints examined (see Table 3 for an example of a typical report). Proficiency and certification tests offered by the profession do not require examiners to show their work, so there is no way to determine the extent to which they followed ACE-V correctly and consistently. Nor are examiners required to document their comparison method steps when they testify to an identification in court. In contrast, adequate report forms that show all the steps of an examiner’s work are required in other countries, such as Canada (Ashbaugh, 2005a). Unless examiners show their work, their results cannot be used as evidence of the validity of the method employed.

Insert Table 3 here

Standardized training. Standardized training programs in the ACE-V method are necessary to ensure that examiners are properly trained, and that the method is uniformly applied. The IAI (2004), TWGFAST (1998), SWGFAST (2006), FBI (Meagher, 2002), and ASCLD (2001) have each provided some guidelines for a training syllabus for the ACE-V method. These recommendations are general, not specific. They are not requirements, and (with the possible exception of the FBI training syllabus, described by Meagher, 2002, but not published), there is no evidence that any of these recommendations are actually written or in practice. The training programs

themselves must include assessment procedures, so it can be determined whether individual trainees or working examiners have learned and use the steps of the method correctly. A formalized training program in the official ACE-V method needs to be designed, including specific goals and their assessment. This program must be adopted and required by the profession.

Most examiners receive the majority of their training on-the-job, without either a formal structure of topics covered or formal assessment of success in meeting training goals. Variation in training means that variation in the use of a method must also occur.

Proficiency assessment. The subjects who serve in the validity experiment must have demonstrated high proficiency in the ACE-V method, so that any errors made can be attributed to the method and not to poor performance by the examiners. To do this, subjects must have specific training on the approved manual, and, most importantly, have demonstrated high skill in using the method. However, present assessment of proficiency of examiners in their training and in their case work is incomplete and inadequate (Haber & Haber, 2004) for several reasons.

For example, proficiency tests and procedures have never been assessed for their validity or their reliability (Haber & Haber, 2004). The validity of a proficiency test would be shown by high correlation with other independent measures of skill and ability, such as supervisor ratings, or the quality and quantity of training and experience. The proficiency test manufacturers have never reported any correlations with these independent measures, so nothing is known about the validity of thee tests. Further, no information has ever been reported on the reliability of these tests, the degree to which examiners receive the same score when they take a comparable form of the test again. If not reliable, they cannot be valid. If the present tests were assessed, it is likely that those presently in use, such as the IAI certification test (IAI, 2004), the IAI-ASCLD proficiency test (Koehler, 1999; Cole, 2005), and the FBI (Meagher, 2002) would fail to exhibit acceptable levels of validity or reliability.

In addition, none of the proficiency tests contains fingerprints of known difficulty, because the profession lacks a quantitative measure of print quality (difficulty). One expert observed that the prints used in the FBI proficiency test are so easy they are a joke (Bayles, 2002).

Further, the prints used in proficiency tests do not reflect normal casework. They are predominately or entirely of value, in contrast to case work, in which the majority of latent prints are of no value. These proficiency tests do not include many, if any, exclusions, though, again, the most common outcome in case work is exclusion. When an examiner receives a particular score on such a test, it is impossible to interpret that score other than relative to other examiners who took the same test. The results cannot be generalized to the examiner's performance on the job, or accuracy in court, because the difficulty of the test items is unknown, and the other parameters do not correspond to normal case work.

Consequently, in the absence of a standardized description of ACE-V, a standardized syllabus and training objectives, and measures of performance during each step of a comparison, it is impossible to assess an examiner's level

of proficiency (accuracy) in the ACE-V method. If the validity experiment is to address the accuracy of the method, highly skilled examiners must perform the comparisons. At present, there is no objective way to select such a group.

What Is the Evidence for ACE-V Validity?

No such experiment has ever been offered as evidence in any of the Daubert hearings, nor has one ever been published. A recent FBI publication (Budowle, et al., 2006), Haber and Haber (2004), and Cole (2005) reported failure to find a single peer reviewed study that tested the validity of ACE-V.

As further evidence of the absence of such an experiment, the National Institute of Justice issued a solicitation for research proposals in 2000 that included testing the validity of the method (NIJ, 2000). That solicitation was never funded, amid controversy (Samuels, 2000; US v. Plaza, 2001). A comparable solicitation was repeated late in 2004 (NIJ, 2004).

A recent experiment by Wertheim, Moenssens, and Langenburg (2006) purported to provide some measure of error rates for examiners using the ACE-V method. The experimenters recorded the accuracy of a large number of latent-to-exemplar comparisons from 108 examiners, collected over several separate week-long training courses. They found that their 92 participants with more than one year of experience made only 81 erroneous individuations out of a total of 6,441: less than a 2% error rate.

We provided a detailed critique of this experiment (Haber & Haber, 2006) in which we argued that the experiment cannot be used to assess experimenter error rate (proficiency). This experiment did not use uniformly well-trained examiners, it was carried out in a training and not a case work environment, the difficulty of the latents was adjusted to the participant's skill level by the course instructor to insure success in this part of the course requirement, the conditions did not match normal working conditions, every latent being compared always had a correct match in the packet, no latent print could correctly be judged of no value, extra help in the form of hints was available from the instructor, and participants could determine on which latents they were to be scored, thereby greatly reducing the risk of reaching a wrong conclusion (partially completed packets could be returned without penalty). Finally, no documentation was required as to what steps the participants followed when making comparisons. Therefore, the results cannot be used to estimate examiner error rate when using the ACE-V method.

III. Evidence of Validity Presented by Fingerprint Examiners

Examiners in Daubert hearings have asserted that ACE-V has been tested, and that the error rate is very low, or even zero (Meagher, 1999). To support these claims, they offer evidence from history, from adversarial testing, from verification testing, from quality of training and experience, from types of errors that could be made, from publicity, from the scientific method, and from proficiency test results. We summarize and critique there these eight claims that the validity of the ACE-V has been demonstrated.

100 Years of Fingerprint History

Many government witnesses have testified that the validity of the ACE-V method has been tested by its continued acceptance throughout 100 years of

history (e, g., Meagher, 2002). However, the hundred years of acceptance provides only "a face" validity which argues that fingerprint individuations must be accurate because people have believed them for a long time. Face validity is not a scientific test of accuracy, only of belief. Further, no one has claimed that the ACE-V method, as distinct from fingerprint evidence, has been accepted for 100 years. As recently as 1988, the FBI's Science of Fingerprints does not refer to the comparison method as ACE-V.

Adversarial Testing

Government witnesses in some Daubert hearings have argued that the results of application of the ACE-V method are tested through the adversarial process during each trial itself (e.g., US v. Havvard, 2001). Since it is claimed in these hearings that no erroneous individuations have ever been uncovered during direct and cross examination, this procedure of testing shows the error rate of the method must be zero.

Adversarial testing does not provide a mechanism to assess the error rate of ACE-V. Ground truth is unknown during adversarial proceedings, so the outcome cannot be used to assess validity of the method being used. Further, the vast majority of cases involving fingerprint evidence result in plea bargains, or the fingerprint evidence goes unchallenged and therefore never are subjected to the adversarial process or to second opinions.

Verification Testing

Some government witnesses have argued that verification procedures represent a testing of the method, since a second examiner checks the conclusion of the first (US v. Havvard, 2001), thereby guaranteeing that errors do not occur.

Verification testing fails in several ways to provide evidence of validity. In case work verification testing, ground truth is unknown, and agreement between two examiners might mean either that they both were correct in the identification, or that they both made an error either by chance or carelessness, or because some property of the method led both to make the error. Further, most verification testing in crime laboratories is non-blind, which permits contamination and bias to reduce the chances of detecting errors (Haber, 2002; and see our discussion under the Standards Criterion below). Crime laboratories closely guard and do not publish results on the number of verifications they do, the number of those that produced different conclusions, how those differences were resolved, and whether the differences are resolved in ways that reduce errors. The extent to which errors are reduced by current practice is simply unknown. Cole (2005) lists a number of instances of erroneous identifications made in court, nearly all of which had been verified by another examiner, each thereby counting as two erroneous identifications. This list is evidence that errors do occur that get past verifiers.

Skilled Examiners do not Make Errors

The government has claimed that erroneous identifications are only made by poorly trained or inexperienced practitioners. When the method is used by well trained and experienced examiners, no errors are ever made, so that the method itself is error-free (Meagher, 2002; 2003).

A number of scientists have noted that if the cause of errors is attributed to practitioners because of their inadequate training and experience, examiner training and experience need to be standardized and tested, and the results of those tests must be known for each examiner before an error occurs. Otherwise, this reasoning is circular (Cole, 2006). The three FBI examiners who concurred in the misidentification of the Madrid bomber (Stacey, 2004) were among the most senior, most experienced, and most trained at the FBI.

Errors involving Value, Exclusion or Inconclusive are not Important

Meagher (2002) has also argued on behalf of the government that the only methodological error rate of relevance to Daubert courts is the percent of erroneous individualizations. Other kinds of errors that can result from the application of ACE-V, including missed identifications based on erroneous exclusions and inconclusive conclusions, are irrelevant and should not be considered in evaluating the validity of ACE-V. If there are no erroneous identifications, then the error rate of the method is zero.

This argument fails to consider that erroneous individuations and erroneous exclusions are necessarily highly correlated. If an examiner wants to avoid the possibility of making erroneous individuations, he can simply make fewer identification conclusions. Doing so obviously results in an increase in the rate of perpetrators missed. If the crime laboratory, profession or courts want to minimize erroneous identifications, they can guarantee that outcome by increasing the punishment to examiners for making such an error. From a scientific standpoint, assessment of a method's error rate must reflect the ratio of correct to incorrect conclusions that result from application of that method.

Publicity

As further evidence of a zero error rate, Meagher (2002) reported that in his 35 years working for the FBI, he had never seen or heard of an erroneous identification made by an FBI agent. Since such an error would be widely publicized, a lack of publicity can assure the court that no such errors had ever occurred.

When FBI Agent Meagher claimed that no erroneous identifications by FBI agents have occurred because he had never heard of one, he failed to consider that the chances of uncovering an erroneous identification are remote. Most fingerprint identifications are not challenged in court, either because the defendant pled to some other charge, or because the defense did not obtain a second opinion. Further, after conviction, the opportunities for innocent persons to obtain new evidence and have their convictions reviewed and overturned are still extremely rare.

ACE-V is based on the Scientific Method

Several prominent examiners (Ashbaugh, 1999; K. Wertheim, 2003), have argued that that because ACE-V involves hypothesis testing by an examiner, which is a component of the scientific method, that comparability means that the ACE-V method itself is as valid as the scientific method is valid.

This argument is flawed. An analogy between the ACE-V method and scientific methods of hypothesis testing does not provide evidence of accuracy. It is an analogy, and nothing more. A method is demonstrably accurate (valid)

when its application consistently produces conclusions that agree with ground truth.

Proficiency Test Results

Evidence from practitioner performance testing (proficiency and certification) is claimed by the government (Cole, 2005) to provide estimates of error rates of the ACE-V method. This argument is strange, because (Haber & Haber, 2004) and Cole (2005) have shown that the published proficiency and certification test results show many errors, including erroneous identifications. We would be the first to argue that because the proficiency tests themselves do not reflect casework and have not been validated, they cannot be used as evidence of examiner accuracy—or inaccuracy.

More importantly, because none of the published tests in use requires the examiner to document the method he used to reach his conclusions, the results of these tests cannot be used evidence for the validity (accuracy) of ACE-V.

None of the eight arguments offered by the government involves scientific tests and none addresses assessment of the validity of the ACE-V method. Cole (2006) provides a detailed examination of the fallacy of these arguments with respect to scientific testing. He concludes, with us, that none of these arguments bears even indirectly on whether the ACE-V method is a valid procedure on which to base a conclusion.

IV. Standards Required for the Validity of a Method

An evaluation of a scientific methodology includes evidence that the conclusions reached from the application of the method rest on quantitative standards that themselves have been shown to be valid. Validity of a standard requires that the standard is objective, tested, and has been calibrated against ground truth. We consider the four standards on which ACE-V rests: value, exclusion, individuation and inconclusive (SWGFAST, 2002a, b). Each of these conclusions rests on a separate standard that can be evaluated for its agreement with ground truth. Validation of these standards involves separate experiments, which themselves require prerequisites before they can be run (see Table 4).

Insert Table 4 here

Analysis Standard: Value

The first standard the examiner applies concerns whether there is sufficient ridge and feature detail in the latent print to attempt a subsequent comparison: whether the latent print is of value. Value is often expressed in terms of the difficulty of the latent print. If the examiner judges the amount of detail in the latent to exceed this value standard, he continues to analyze the latent fingerprint. If he judges the amount of detail insufficient to meet the value standard, he desists from further work on that latent.

Demonstration of the validity of the value standard requires experimental evidence that highly skilled examiners agree on whether a specific latent print is of value for comparison. If the value standard is to rest on a quantitative basis, as it must for the standard itself to be quantitative, it requires three prerequisites: a quantitative metric of the difficulty of latent prints; evidence that examiners (or computer algorithms) can evaluate the difficulty of individual latents consistently; and a quantitative statement by the profession that latent prints that fail to exceed

a specific threshold of difficulty are deemed of no value for comparison (see Table 4). We consider these three prerequisites briefly.

A quantitative metric of difficulty requires an experiment. A large number of latent fingerprints are randomly selected from case work. Random selection insures that the proportion of latents of no value in this sample of latents reflects the typical number found in casework. Skilled examiners are then asked to evaluate each latent print in two ways. First, they are to judge the overall difficulty of the latent on a rating scale. Second, they evaluate each print on separate rating scales for a number of specific dimensions expected to predict difficulty. These include, for example, overall size, level one classification, presence of focal features such as a core or delta, presence of orientation information, contrast, smear, amount of distortion, level two features such as ridges that can be followed or the presence of minutiae on the ridges, and level three features such as pores and ridge edges. Analysis of the rating scale scores across the latent prints would establish the different dimensions of difficulty, so that a total difficulty score could be calculated for each latent, a score that is weighted for the relative importance of each rating scale determined by multivariate analysis of the rating scales. From these data, it is possible to rate objectively the difficulty of any new latent fingerprint on these scales.

The validity of the use of these latent print difficulty scales can then be demonstrated by an experiment. A random sample of examiners can be given a selection of latent prints ranging in rated difficulty (as described above). If examiners can use these scales accurately, they would agree on the presence of the difficulty variables, and the overall difficulty level of each latent. Once such a metric is established, computer programs could be created so that latent print difficulty could be assessed speedily and uniformly. Yao, et al. (2004) provide an initial attempt to create such a program.

No experiment to quantify latent print difficulty has ever been published. No experiment has been performed to determine the consistency with which examiners judge a particular latent print to be of value. However, two kinds of available evidence suggest examiners differ in their value judgments.

First, skilled examiners assert (Ashbaugh, 1999, 2005b; K. Wertheim, 2003) that the value of a latent depends on the training and experience of the examiner, so that one examiner might judge a latent as of no value, whereas another might be willing to proceed to compare it and trust his conclusions. If the standard is training-and-experience dependent, then definitions of difficulty AND definitions of training and experience are each required before an objective value standard can be validated.

Second, many of the proficiency tests administered for the IAI and for crime laboratories have required the examiners taking the test to indicate the value of the latent prints. The reported data show that the variation (error rate) on that conclusion is quite high on those tests (Haber & Haber, 2004).

Comparison Standard: One-Unexplained-Discrepancy

During Comparison, the examiner may conclude an exclusion of a suspect's exemplar digit(s) on the basis of the one-unexplained discrepancy standard (Thornton, 1977). This standard is critical to the profession because its

absence or violation abrogates the uniqueness principle underlying forensic comparison.

Demonstration of the validity of the one discrepancy standard requires experimental evidence that examiners differentiate accurately between two prints that arise from distortions only (a known ground truth of a single donor) and genuine discrepancies (a known ground truth of two donors). No experimental evidence has been published on the one unexplained discrepancy standard.

An experiment to demonstrate the validity of the one discrepancy standard would score examiners on the accuracy with which they apply this standard. The test materials would consist of a set of exemplar-latent pairs, for which all latents are of value, and the ground truth of each pair is known. Starting with the first pair, each examiner would first do a complete analysis of the latent print. The examiner documents a complete record of the analysis (including, for this experiment, special attention to distortions) on a detailed report form. When the examiner completes the analysis by selecting an area of the latent print to use for the beginning of comparison, the same portion of the exemplar print is provided, with the rest of the exemplar print masked off. The examiner's task is to describe the amount of agreement found, note any discrepancies, and for each discrepancy found, account for its cause from the analysis of the distortions in the latent he has already provided. If any of these discrepancies are inexplicable on the basis of the distortions, he should conclude elimination. If every discrepancy can be accounted for by distortions in the latent print, the examiner should put that pair aside, and go onto the next latent print. If examiners are able to use the one-unexplained discrepancy standard, then every exclusion conclusion should be matched with a ground truth of different donors.

However, this experiment cannot be carried out until the following prerequisites are met (Table 4). The profession needs to write and adopt a report form for the Analysis stage which includes a checklist of distortions potentially present in every latent, and a formalized mode of describing the grounds on which each distortion was judged to be explicable or inexplicable during Comparison. Until this report form is available, the validity of the one-unexplained discrepancy standard cannot be demonstrated.

Indirect evidence suggests that examiners differ in their judgments regarding an unexplained discrepancy. Many examiners begin Analysis by looking at the exemplar print and latent print side by side, or, during Comparison, looking back at the latent for features newly discovered in the exemplar. This practice lets the examiner "reshape" the latent. When the analysis of the latent is contaminated by knowledge of the patterning in the exemplar, the one-discrepancy standard is open to errors of bias (Haber & Haber, 2005). Using features from the exemplar to identify ones in the latent creates opportunities for the examiner to overlook inexplicable discrepancies (Ashbaugh, 1999; 2005b).

Evaluation Standard: Numeric Sufficiency

During Evaluation, the examiner may conclude an individuation or inconclusive, on the basis of the number of features in agreement between the two fingerprints. If the amount exceeds the numeric sufficiency standard, the examiner concludes that the two fingerprints have the same donor. If the amount

of agreement fails to meet the sufficiency threshold, he concludes that the suspect's finger can neither be included nor excluded as the donor of the latent print.

Demonstration of the validity of the sufficiency standard requires evidence that the conclusions, based on the standard, are consistent with ground truth. Before this experiment can be run, four prerequisites must be met (Table 4). First, the profession must specify the features found in latent fingerprints that are to be counted. Second, the profession needs to specify how the location of each feature in the latent is defined, either in absolute locations, or relative to every other countable feature. Third, the profession needs to specify rules that determine what constitutes agreement of features by kind and location between latent and exemplar, and what constitutes non-agreement. Fourth, evidence is needed to determine whether examiners apply these rules consistently.

Once these prerequisites are fulfilled, then an experiment to demonstrate the validity of the numeric sufficiency standard can be carried out. The purpose of the experiment is to establish a quantitative metric of agreement between latent and exemplar, with a concomitant probability of matching ground truth. Skilled examiners are given a large number of latent prints, all of value but of varying difficulty. Each latent print is to be fully analyzed. Then, each latent print is paired with an exemplar, some of same donor ground truth, some of different donor, and the pair is to be compared and evaluated. The examiner records every corresponding point in agreement by type and location between the two prints.

It is to be expected that as the number of features in agreement increases, the probability of ground truth being a single donor increases. For example, it might be found that when ten points are found in agreement, 80% of the pairs have a ground truth of one donor, and 20% of the pairs with ten points in agreement have two donors. It might be further found that 95% of the pairs with 25 points in agreement have a ground truth of one donor, with only 5% having two donors. The profession could then decide what value to choose for the sufficiency standard, depending on the percentage of erroneous identifications the profession is willing to accept when the examiner concludes identification.

The available evidence suggests that skilled examiners vary in the number of features they perceive in a latent, and in the number of features they count as in agreement in latent and exemplar. Langenburg (2004) reported that even skilled examiners differed as to the number of minutiae they labeled in latent prints. Evett and Williams (1996) reported data showing substantial variation in judgments from highly experienced examiners in the amount of agreement between each of 10 latent-exemplar pairs (all of value). On one pair, judgments ranged from 14 to 51 points of agreement. The Langenburg and Evett and Williams results show the critical importance to the profession of specifying what constitutes a feature, and what constitutes agreement. The current variability among examiners means that a quantitative test of a sufficiency standard based on number of features in agreement will fail to meet reasonable standards of validity.

Until 1973, examiners in the United States based their conclusions on a numeric sufficiency standard. In 1973 the IAI ruled that all numeric standards based on a point count of correspondences be abandoned, because there was no scientific evidence to support any particular numeric standard (IAI, 1973; Stacey, 2005; Budowle, et al., 2006). No experiment such as the one we describe has been carried out.

Evaluation Standard: Training and Experience

In 1973, the numeric sufficiency standard was replaced by a subjective, personal "training and experience" sufficiency standard (Ashbaugh, 2005b). The examiner asserts the equivalent of the following under oath: "Based on my training and my experience, I have never seen this amount of agreement between two fingerprints obtained from two different people—therefore I am absolutely confident that this amount of agreement means that the same person made both fingerprints" (wording suggested by Ashbaugh, 2005b). However, when an examiner refers to his own experience as the source of the standard, he forgets that his experience (case work) has no access to ground truth. Hence, there is no way for him to determine whether some amount of agreement is an identification grounded in truth.

An experiment to test the validity of a training and experience sufficiency standard has several prerequisites (Table 4). Training needs to be specified in quantitative units, and experience needs to be specified in quantitative units. Neither of these exists today.

Given that current training is not standardized, and that experience is heterogeneous, and that the profession has not grappled with defining either training or experience quantitatively, neither of those definitions can be specified today. Therefore, the subjective personal sufficiency standard based on training and experience in its present form cannot be validated.

Once training and experience are quantitatively defined, an experiment to demonstrate the validity of this sufficiency standard could be performed. Examiners of varying training and experience are asked to compare pairs of latent and exemplar prints of known ground truth, representative of case work in difficulty. The presumed results are that as training and experience increase, so do agreement with ground truth. If sufficient accuracy were found, the profession could then adopt a training and experience sufficiency criterion: a given amount of quantified training and quantified experience assures resulting conclusions with a known probability of being correct.

Because practicing examiners differ greatly in the amount and kinds of training and experiences they have undergone, it is expected that different examiners will reach different conclusions when comparing the same set of prints. There is ample evidence that this is in fact the case, both among highly trained and experienced examiners, and across the entire spectrum of training and experience (Evett & Williams, 1996; the Mitchell FBI survey reported in US v. Mitchell, 2000; and certification examination results, Grieve, 1996). Further, neither the court nor the examiner himself has access to any measure of ground truth on which to estimate the accuracy of the examiner's conclusion. It is simply his opinion, based on his personal experience and training. This subjective and

opinion-based standard differs from a numerical standard, in that with the latter, the validity of the standard can be tested. With the former, which is now the standard required by the IAI, the probability of error is untested, and cannot be tested in current practice.

Verification Standard: Blind Replication

The final standard is stipulated by the fingerprint profession itself, and is under the control of the crime laboratory, not the individual examiner. During Verification of an individuation conclusion, a second examiner compares the latent and exemplar prints. The blind component of the standard (Smrz, et al., 2006; Stacey, 2005; SWGFAST, 2006) requires that in the absence of knowledge of any prior conclusion or who made it, the verifying examiner must reach the same conclusion as the first examiner. However, current practice in nearly all crime laboratories in which verification is performed is non-blind.

Validation of the verification standard requires evidence of the accuracy of the replication process: the percent of errors made by the first examiner that are uncovered by the verifier. Research on quality control (Arvizo, 2002, 2003; Boone, et al., 1982) shows that non-blind verification catches relatively few errors, whereas blind verification, and especially double blind verification, catches many more errors. Blind verification is standard procedure in governmental testing laboratories (for example, FDA Guidelines, 2001), and is a requirement for publication in virtually all scientific journals.

Fingerprint examiners have testified for the government that non-blind verification catches most errors (Meagher, 2002). This claim can be tested in an experiment by comparing non-blind with single blind procedures. “Problem” pairs of latent-exemplar prints known to have resulted in mistakes in training, on proficiency tests, or in court are given to a number of examiners who are asked to complete an ACE-V procedure and offer a conclusion. Ground truth for these pairs may not be known, but the selection of pairs is restricted to those for which expert examiners have consistently agreed which of the pairs are identifications and which exclusions. Half of the pairs (of both identifications and exclusions) are distributed among the examiners with the notation that this pair has already been individuated (or excluded) by a skilled examiner and they are to confirm the conclusion (non-blind). The other half of the pairs (of both identifications and exclusions) is distributed among the examiners as if each is a new pair, and each examiner is asked to use ACE-V to reach a conclusion (blind). The results would compare, between the non-blind and blind conditions, the number of times the examiners differed from the experts’ classification of the pair.

This experiment has never been reported in this form. However, Dror and his colleagues have reported several experiments to show that skilled examiners can be induced to change their initial conclusion after presentation of biasing information. Dror, Charlton and Peron (2006) selected five senior examiners (who had previously volunteered to be tested from time to time for research purposes) and asked them to do a complete ACE-V on a single latent-exemplar pair. The pair for each examiner was selected from that examiner’s file (unbeknownst to the examiner), in which the examiner had made an identification conclusion at least five years earlier. The pair in each case was re-presented to

the examiners as an example of the erroneous identification of Brandon Mayfield made by the FBI (Stacey, 2004). While the Mayfield case had just broken and the five examiners were familiar with the facts, none of them had yet seen the Mayfield latent or exemplar print and none remembered the pair they had personally individuated years earlier. One of the 5 examiners changed his previous identification conclusion to inconclusive, and 3 changed their previous identification to exclusion. Only one verified his previous identification. We equate the biasing information presented in this experiment with non-blind replications: the examiner has information about the expected outcome of his work.

In a follow up experiment, Dror and Charlton (2006) tested six examiners on four pairs they had previously concluded identification and four pairs they had previously excluded. When the 48 pairs were now presented a second time (again, unbeknownst to the examiners), half were accompanied by biasing information (for the earlier exclusions, the examiners were now told the suspect confessed; for the earlier identifications, they were told that the suspect had been in jail at the time of the crime). For the 24 pairs on which bias was introduced, four instances of identification were changed to exclusion. No exclusions were changed to identification. For the 24 pairs on which no bias was introduced, one examiner changed an identification to an exclusion conclusion, and one changed an exclusion to an identification. This finding shows variability in accuracy even in the absence of bias.

Dror's work shows that non-blind replication led to change consistent with the introduced bias. Even more troubling in the context of the validity of ACE-V, Dror's data show that highly skilled examiners reach different conclusions with themselves over time, even in blind replications of their own work. See also Dror, Peron, Hind & Charlton, 2005) for another example, and Dror & Rosenthal (2006) for further statistical analyses of the experiments discussed here.

Evidence Presented by the Government on the Validity of the Four Standards

The Daubert ruling does not refer explicitly to validation of the standards inherent in a method. We, as research scientists, have described the kinds of evidence needed to demonstrate the validity of the four conclusions a fingerprint examiner may draw from application of the ACE-V method. Fingerprint examiners, testifying on behalf of the government in Daubert hearings, have not addressed them, other than negatively to note that the numeric sufficiency standard was abandoned in 1973 due to lack of validation.

V. Which Community of Experts: Examiners or Researchers?

The 40+ courts in which the validity of fingerprint evidence was challenged have unanimously ruled that this evidence was admissible. This decision reflects the testimony of fingerprint examiners, who affirm that the ACE-V method is valid, and that fingerprint evidence has been accepted for one hundred years.

One of the criteria used under Daubert concerns evidence that the members of relevant scientific communities accept the method as valid (Saks, 2003; Faigman, et al., 1997, 2002). Fingerprint examiners certainly comprise a relevant community. However, with rare exceptions, they are untrained in

experimental science and lack the qualifications to evaluate the validity of a method.

When other scientifically based methods have undergone the scrutiny of a Daubert hearing, most, if not all of the testimony has rested on evidence presented by scientists trained to evaluate the research evidence of the validity (or non-validity) of the method in question. These scientists all belong to a community of researchers who have the training and skills to carry out and publish the empirical studies that demonstrate the method's validity.

However, in the Daubert hearings on the acceptance of the validity of the ACE-V method, fingerprint examiners argued that the fingerprint profession is qualified to speak to the validity of the method they use, and that only they are qualified to provide the scientific evidence of the validity of the method used to make the forensic comparisons (See Cole, 2005, for a review of these claims). The fingerprint community is the scientific group intended under Daubert (Meagher, 2002). In some Daubert hearings, the dissenting testimony by scientists has been ruled inadmissible (US vs. Havvard, 2001).

The uniform rulings to accept fingerprint examiners as the relevant scientific community is surprising, given the 1993 Daubert Court ruling. Furthermore, in accepting practitioners as the relevant community, these courts have failed to recognize a dichotomy in the backgrounds and training of the experts providing the testimony in the hearings. For example, no research scientist familiar with comparative procedures underlying ACE-V has ever testified in a Daubert hearing or submitted an Amicus brief on the government side. No research scientist who has testified in a Daubert hearing on the validity of the ACE-V method has accepted that the validity of the ACE-V method has been demonstrated.

As a different example, legal scholars familiar with the scientific evidence on the validity of ACE-V have never testified or submitted an amicus brief in a Daubert hearing on the government side. When legal scholars have testified in Daubert hearings regarding fingerprint evidence, they have been almost unanimous in attesting to the absence of scientific validity evidence of the ACE-V method. A substantial legal scholarship published literature attacks the fingerprint profession for its failure to test the validity of the ACE-V method (Saks, 2003; Faigman, et al, 1997; 2002).

As a final example of this dichotomy, in a recent Amicus Brief submitted to the Massachusetts Supreme Judicial Council (Massachusetts v. Patterson, 2005), five Ph.D. scientists and 12 legal scholars signed the brief opposing the admission of fingerprint evidence based on problems with its validity. In the same case, the government briefs were contributed only by fingerprint examiners or prosecutors, and did not include a single contribution from a research scientist or a legal scholar.

Fingerprint evidence has a 100 year history of court acceptance, but the ACE-V has not been systematically tested for validity. Belief in its accuracy by the professionals who practice the method does not constitute evidence of acceptance by a relevant scientific community.

VI. Conclusion

We have reviewed available scientific evidence of the validity of the ACE-V method and found none. However, we report a range of existing evidence that suggests that examiners differ at each stage of the method in the conclusions they reach. To the extent that they differ, some conclusions are invalid.

We have analyzed the ACE-V method itself, as it is described in the literature. We found that these descriptions differ, no single protocol has been officially accepted by the profession, and the standards upon which the method's conclusions rest have not been specified quantitatively. As a consequence, at this time the validity of the ACE-V method cannot be tested.

We laid out in some detail the kinds of experiments needed to assess the method's validity, including the prerequisite research to create objective measures. The prerequisites for a test of the validity of the ACE-V method were presented in Table 2. The prerequisites to test the validity of the standards inherent in ACE-V were presented in Table 4.

More than 40 Daubert courts have ruled to admit fingerprint comparison evidence because it meets validity requirements. We view Tables 2 and 4 as a refutation of the claim that there is scientific evidence for the validity of ACE-V. These tables show that both fingerprint examiners and researchers need to do a substantial amount of work to quantify, test, train and validate the components of ACE-V. Only then can the validity of the ACE-V method be determined.

Comparison conclusions have been accepted by US courts for one hundred years, and fingerprint examiners believe their conclusions are accurate. Neither of these beliefs documents the scientific validity of the method: evidence that the conclusions reached by proper application of the method agree with ground truth, and with a known probability of error.

As research scientists, we are deeply concerned by the Daubert courts' decisions to admit fingerprint evidence. The consequence: fingerprint individuations based on a method of unknown validity are being used to convict suspects, an unknown number of whom are innocent.

References

American Society of Crime Laboratory Directors (2001). Laboratory accreditation program. http://www.ascld-lab.org.

Altschuler, A. (2005) Quoted on Public Television's program ***The Plea***, produced by Ofra Bikel.

Ashbaugh, D. (1999). Quantitative-qualitative friction ridge analysis: An introduction to basic and advanced ridgeology. Boca Ratan, Fl. CPC Press.

Ashbaugh, D. (2005a). Friction Ridge Identification process ACE-V worksheet. Available at http://www.furrowsandridges.homestead.com/A_C_E_-V_worksheet.pdf.

Ashbaugh, D. (2005b). Class notes from IAI sponsored training course on latent print comparison.

Arvizu, J. (2002). Testimony in US v. Plaza,188, F Supp.22nd, 2002 Daubert hearing.

Arvizu, J. (2003). Speaker at the Uses and Misuses of Forensic Evidence Symposium, March 6, 2003.

Bayles, A. (2002). Testimony in US vs. Plaza, 188, R. Suppl., 2d, 2002 Daubert hearing.

Beeton, M. (2001) Scientific methodology and the friction ridge identification process. http://ridgesandfurrows.homestead.com/files/scientific_methodology.pdf.

Boone, D. J., Hansen, H.J., Hearn, T.L., Lewis, D.S., & Dudley, D. (1982). Laboratory evaluation and assistance efforts: mailed, on-site, and blind proficiency testing surveys conducted by the Centers for Disease control. American Journal of Public Health, 72, 1364-1368.

Budowle, B., Buscaglia, J., & Perlman, R.S. (2006). Review of scientific basis for fingerprint comparisons as a means of identifications: committee findings and recommendations. Forensic Science Communications, 8, 1-16 (www.fbi.gov/hg/law/fsc/current/research/2006)

Champod, C., Lennard, C. Margot, P., & Stoilovic, M. (2004). Fingerprint and other friction ridge skin impressions. Boca Ratan, Fl., CPC press.

Cole, S. (2001). Suspect Identities: A history of fingerprinting and criminal identification. Cambridge: Harvard University Press

Cole, S. (2004). Grandfathering evidence: fingerprint admissibility rules from Jennings to Llera Plaza and back again. American Criminal Law Review, 41, 1189-1276.

Cole, S. (2005). More than zero: Accounting for error in latent print identifications. Journal of Criminal Law and Criminology, 95, 985-1078.

Cole, S. (2006) Is fingerprint identification valid: Rhetoric of reliability in fingerprint proponents' discourse. Law and Policy, 28, 109-135.

Daubert v. Merrell Dow Pharmaceuticals, Inc. 509, US, 579 (1993)

Dror, I.E., & Charlton, D. (2006). Why experts make errors. Journal for Forensic Identification, 56, 600-616.

Dror, I.E., Charlton, D., & Peron, A.,E. (2006) Contextual ;information renders experts vulnerable to making erroneous identifications. Forensic Science International, 156, 74-78.

Dror, I.E., Peron, A., Hind, S., & Charlton, D. (2005). When emotions get the better of us: The effects of contextual top-down processing on matching fingerprints. Journal of Applied Cognitive Psychology, 19, 799-809.

Dror, I.E., & Rosenthal, R. (2007). Meta-analytically quantifying the reliability and bias ability of fingerprint experts' decision making. Southampton University Technical Report.

Epstein, R. (2002). Fingerprints meet Daubert: the myth of fingerprint 'science' is revealed. Southern California Law Review, 78, 605-625.

Evett, I.W., & Williams, R.L. (1996) A review of the 16 point fingerprint standard in England and Wales. Journal of Forensic Identification, 46, 49-73.

Faigman, D.L., Kaye, D.H., Saks, M.J., & Sanders, J. (1997; revised 2002). Modern Science Evidence: the Law and Science of Expert Testimony. St. Paul, Minn.: West Publishing.

FBI Science of Fingerprints (1988). Revised edition; first edition, 1958. Washington: US Government Printing Office.

Food and Drug Administration. (2001). Guidance for Industry: Choice of control group and related issues in clinical trials. 29 pp.

Grieve, D. (1996). Possession of truth. Journal of Forensic Identification, 46, 521-528.

Haber, R.N. (2002). Testimony in US v. Plaza, 188, F Supp. 2d, 2002 Daubert hearing

Haber, L., & Haber, R.N. (2004). Error rates for human latent fingerprint examiners. In N. Ratha & R. Bolle, (Eds.). Automatic fingerprint recognition systems (pp 339-360). New York: Springer

Haber, L., & Haber, R.N. (2005). Mindset in the latent print comparison process. Lecture presented to the IAI Annual Educational Conference, Dallas, TX, August, 2005.

Haber, L., & Haber, R.N. (2006). Re: A report of latent print examiner accuracy during comparison training exercises. Journal for Forensic Identification, 56, 493-499.

Huber, R. (1959). Expert witness. Criminal Law Quarterly, 2, 276-297.

Huber, R.A. (1972). The philosophy of identification. Royal Canadian Mounted Police Gazette, July/August, 1972.

International Association for Identification (2004). Latent print examiner certification requirements: subsection on training requirements. http://www.theiai.org/certifications/fingerprint/requireements.html.

International Association for Identification (1973). Report of the Standardization Committee, Journal of Forensic Identification,

Interpol European Expert Group on Fingerprint Identification (IEEGFI II) (2005). Method for Fingerprint Identification. http://interpol.int/public/forensic/fingerprints/workingparties

Koehler, J.J. (1999). Review of Collaborative Testing Services Proficiency Tests. Technical Report, McComb School of Business and School of Law, University of Texas at Austin.

Langenburg, G.M. (2004). Pilot study: A statistical analysis of the ACE-V method—analysis stage. Journal of Forensic Identification, 54, 64-79.

Legal challenges to fingerprints. FBI Onin Website. http://onin.com/fp/daubert (updated Sept., 2005)

Massachusetts v. Terry Patterson (2005) Reliability of latent print identification . Brief of Amici Curiae on behalf of scientists and scholars by the new England Innocence Project. Massachusetts Supreme Judicial Council, SJC 09478).

Meagher, S. (1999). Testimony in US v. Mitchell, N. 96-4071 (ED PA, 1999.

Meagher, S. (2002). Testimony in US v. Plaza, 188, F Supp. 2d E. D. PA, Daubert hearing.

Meagher, S. (2003). Interview on CBS 60 Minutes on fingerprint accuracy, January 5, 2003.

National Institute of Justice. (2000). Solicitation for Forensic Friction Ridge (Fingerprint) Examination Validity Studies. Washington, D.C.: United States Department of Justice.

National Institute of Justice (2004). Solicitation on Forensic Friction Ridge Fingerprint Examination Validity Studies. Washington, D.C.: United States Department of Justice.

Olsen, R., & Lee, H. (2001). Identification of latent prints. In H.C. Lee & R.E. Gaensslen (Eds.), 2nd ed., pp. 41-61. Boca Ratan, FL: CRC Press.

Saks, M. (2003). Reliability standards: too high, too low, or just right: The legal and scientific evaluation of forensic science (especially fingerprint expert testimony. Seaton Hall Law Review, 33, 1167-1187.

Samuels, J (2000). Letter from NIJ Regarding the Solicitation for Friction Ridge (Fingerprint) Examination Validation Studies, Forensic Science Communications, 2, pp. 1-2.

Smrz, M.A., Burmeister, S.G., Einseln, A., Fisher, C.L., Fram, R., Stacey, R.B., Theisen, C.E., and Budowle, B (2006). Review of FBI latent print unit processes and recommendations to improve practices and quality. Journal for Forensic Identification, 56, 402-434.

Stacey, R. (2004). A report on the erroneous fingerprint individualization in the Madrid bombing case. Journal of Forensic Identification, 54, 706-718

Stacey, R. (2005) A report on the erroneous fingerprint individualization in the Madrid bombing case. Presented at the IAI Education Conference, Dallas, August 9, 2005.

SWGFAST (2002a). Friction ridge methodology for latent print examination. http://www.swgfast.org/F_R_M_L_P_E/1.01.pdf

SWGFAST (2002b) Friction Ridge Examination Methodology for latent Print examiners. Available at www.swgfast.org/friction_ridge_examination_print_examiners_1.01.pdf+swgfast

SWGFAST (2006) Quality assurance guidelines for latent print examination. Journal of Forensic Identification, 56, 117-128.

Technical Working Group for Friction Ridge Analysis, Study and Technology (TWGFAST), 1998). Minimum qualifications for latent print examiners trainees. Journal of Forensic Identification, 48.

Thornton, J.I. (1977). The one dissimilarity doctrine in fingerprint identification. International Criminal Police Review, 32, (306), 89-95.

Triplett, M., & Cooney, L. (2006). The etiology of ACE-V and its proper use: An exploration of the relationship between ACE-V and the scientific method of hypothesis testing. Journal of Forensic Identification, 56, 345-356.

Washington v. David Kunze (1999). Court of Appeals for Washington, Div 2, 97, Wash.app.823, 998,P. 2d.

Wertheim, K . (2003). Class notes for training course in fingerprint comparison methods. Santa Ana, CA, December 1-5, 2003.

Wertheim, K., Langenburg, G.M., & Moesssens, A. (2006). A report of latent print examiner accuracy during comparison training exercises. Journal of Forensic Identification, 56, 55-93.

Wertheim, P A. (2002). Scientific comparison and identification of fingerprint evidence. The Print, 16,

US v. Byron Mitchell, N. 96-4071 (ED PA, Feb 2000)

US v. Plaza, Daubert hearing #1, 2001.

US v. Havvard 117 F. Supp 2d, 848,854 (SD Ind.) 2001

US v. Scheffer, 1998, US Supreme Court, 96-1133.

Yao, M., Pankanti, S., & Haas, N. (2004). Fingerprint quality assessment. In N. Ratha & R Bolle (Eds.), Automatic Fingerprint Recognition Systems, pp. 55-66. New York, NY: Springer-Verlag.

Table 1
Classification of ACE-V Conclusions

	Ground Truth	
	Same Donor	Different Donor
Examiner's Conclusion		
No value	missed individuation	missed exclusion
Value + Inconclusive	missed individuation	missed exclusion
Value + Exclusion	erroneous exclusion	correct exclusion
Value + Individuation	correct individuation	erroneous individuation

Table 2
Prerequisites to Test the Validity of the ACE-V Method
1. Profession writes and adopts a detailed ACE-V manual
2. Profession writes and adopts a detailed report form for ACE-V comparisons
3. Profession adopts an approved training program
4. Profession adopts a validated and reliable proficiency measurement

Table 3
Sample Fingerprint Examiner's Report (factual information altered)

Chester Police Department. Crime Laboratory Report
Investigating Officer: John Smith, CLPE
Date of Report: May 1, 2004: Laboratory Number 17-30-4299 LPE
Type of Exam: Latent Print Comparison
Incident Robbery; Location: 1000 block of South Pike Street, Chester

Suspects: Jones, Peter, WM, 9/11/50; CPD Ident # 89765
Evidence: Ten latent print cards, each with one lift; ten-print card of Peter Jones.
Results of Examination:

Each latent print was compared to each finger on ten print of Jones. Eight latents did not match any of Jones' fingers.

Latent #07, taken from inside edge of window sill at point of entry identified to the left ring finger of Jones

Latent # 08, taken from inside edge of counter top next to window sill identified to left index finger of Jones.

Disposition: latent evidence was returned to latent evidence locker and secured.

Table 4. Prerequisites to Validate Standards Underlying ACE-V

Analysis: Value Standard

1. Quantitative metric of latent print difficulty
2. Evidence that examiners use metric consistently
3. Profession adopts a standard of value based on difficulty

Comparison: One Unexplained Discrepancy Standard

1. Profession writes and adopts a report form

Evaluation: Numeric Sufficiency Standard

1. Profession defines features to be counted
2. Profession defines how to specify relative locations among features
3. Profession defines what constitutes agreement
4. Evidence examiners evaluate agreement consistently

Evaluation: Training and Experience Sufficiency Standard

1. Profession defines training in quantitative terms
2. Profession defines experience in quantitative terms

Acknowledgments

We gratefully acknowledge the very helpful comments provided by William Thompson, David Siegel, and Simon Cole.