Update on the State of the Science of Digital Evidence Examination

Fred Cohen – CEO – Fred Cohen & Associates, President – California Sciences Institute

Abstract—*This paper updates previous work on the level of consensus in foundational elements of digital evidence examination. Significant consensus is found present only after definitions are made explicit, suggesting that, while there is a scientific agreement around some of the basic notions identified, the use of a common language is lacking.*

Keywords: Digital forensics examination, terminology, scientific methodology, testability, validation, classification, scientific consensus.

1 Introduction and Background

There have been increasing calls for scientific approaches and formal methods, (e.g., [1][2][3][4][5][6]), and at least one study has shown that, in the relatively mature area of evidence collection, there is a lack of agreement among and between the technical and legal community about what constitutes proper process. [7] The National Institute of Standards and Technology has performed testing on limited sorts of tools used in digital forensics, including substantial efforts related to evidence collection technologies, and it has found that the tools have substantial limitations about which the user and examiner must be aware if reliable tool usage and results are to be assured. [8]

In an earlier paper seeking to understanding the state of the science in digital evidence examination (i.e., analysis, interpretation, attribution, reconstruction, and aspects of presentation),[26] results suggested a lack of consensus and a lack of common language usage. A major question remained as to whether the lack of consensus stemmed from the language differences or the lack of a common body of agreed-upon knowledge. This paper updates the results from that previous work by using a survey to try to differentiate between these two possibilities. In the context of the legal mandates of the US Federal Rules of Evidence [9] and relevant case law, this helps to clarify the extent to which expert testimony may be relied upon.

1.1 The rules and rulings of the courts

The US Federal Rules of Evidence (FRE) [9], rulings in the Daubert case[10], and in the Frye case [11], express the most commonly applied standards with respect to issues of expert witnesses (FRE Rules 701-706). Digital forensic evidence is normally introduced by expert witnesses except in cases where non-experts can bring clarity to non-scientific issues by stating what they observed or did.

According to the FRE [9], only expert witnesses can address issues based on scientific, technical, or other specialized knowledge. A witness qualified as an expert by knowledge, skill, experience, training, or education, may testify in the form of an opinion or otherwise, if (1) the testimony is based on sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case. If facts are reasonably relied upon by experts in forming opinions or inferences, the facts need not be admissible for the opinion or inference to be admitted; however, the expert may in any event be required to disclose the underlying facts or data on cross-examination.

The Daubert standard [10] essentially allows the use of accepted methods of analysis that reliably and accurately reflect the data they rely on. The Frye standard [11] is basically: (1) whether or not the findings presented are generally accepted within the relevant field; and (2) whether they are beyond the general knowledge of the jurors. In both cases, there is a fundamental reliance on scientific methodology properly applied.

The requirements for the use of scientific evidence through expert opinion in the United States and throughout the world are based on principles and specific rulings that dictate, in essence, that the evidence be (1) beyond the normal knowledge of non-experts, (2) based on a scientific methodology that is testable, (3) characterized in specific terms with regard to reliability and rates of error, (4) that the tools used be properly tested and calibrated, and (5) that the scientific methodology is properly applied by the expert as demonstrated by the information provided by the expert.[9][10][11][12]

Failures to meet these requirements are, in some cases, spectacular. For example, in the Madrid bombing case, where the US FBI declared that a fingerprint from the scene demonstrated the presence of an Oregon attorney. However, that attorney, after having been arrested, was clearly demonstrated to have been on the other side of the world at the time in question. [13] The side effect is that fingerprints are now challenged as scientific evidence around the World. [24]

1.2 The foundations of science

Science is based on the notion of testability. In particular, and without limit, a scientific theory must be testable in the sense that an independent individual who is reasonably skilled in the relevant arts should be able to test the theory by performing experiments that, if they produced certain outcomes, would refute the theory. Once refuted, such a theory is no longer considered a valid scientific theory, and must be abandoned, hopefully in favor of a different theory that meets the evidence, at least in circumstances where the refutation applies. A statement about a universal principle can be disproven by a single refutation, but any number of confirmations can not prove it to be universally true. [14]

In order to make scientific statements regarding digital evidence, there are some deeper issues that may have to be addressed. In particular, there has to be some underlying common language that allows the scientists to communicate both the theories and experiments, a defined and agreed upon set of methods for carrying out experiments and interpreting their outcomes (i.e., a methodology), and a predefined set of outcomes with a standard way of interpreting them (i.e., a system of measurement) against which to measure tests. These ultimately have to come to be accepted in the scientific community as a consensus.

One way to test for science is to examine peer reviewed literature to determine if these things are present. This was undertaken in a 2011 study [26] which suggested a lack of common language and a subsequent proposal to move toward a common language based on archival science.[27] One way to test for consensus is to poll individuals actively participating in a field (e.g., those who testify as expert witnesses and authors publishing in relevant peer reviewed publications) regarding their understandings to see whether and to what extent there is a consensus in that community. This method is used across fields [15][16][17], with >86% agreement and <5% disagreement for climatologist consensus regarding the question “Do you think human activity is a significant contributing factor in changing mean global temperatures?” in one survey. [18]

1.3 The previous study being updated

In the previous study of consensus in digital forensics evidence examination, which we quote liberally from with permission in this section,[25] results suggested a lack of consensus surrounding a series of basic statements:

1 Digital Evidence consists only of sequences of bits.
2 The physics of digital information is different from that of the physical world.
3 Digital evidence is finite in granularity in both space and time.
4 It is possible to observe digital information without altering it.
5 It is possible to duplicate digital information without removing it.
6 Digital evidence is trace evidence.
7 Digital evidence is not transfer evidence.

8 Digital evidence is latent in nature.
9 Computational complexity limits digital forensic analysis.
10 Theories of digital forensic evidence form a physics.
11 The fundamental theorem of digital forensics is "What is inconsistent is not true".

These statements were evaluated by survey participants against a scale of “I disagree.”, “I don't know.”, and “I agree.”, and participants were solicited from the members of the Digital Forensics Certification Board (DFCB), individuals who have authored or co-authored a paper or attended the International Federation of Information Processing (IFIP) working group 11.9 (digital forensics) conference over the last three years in Kyoto, Orlando, and Hong Kong, members of a Bay Area chapter of the High Tech Crime Investigators Association (HTCIA), and a group of largely university researchers at an NSF-sponsored event. Control questions were used to control for random guessing and consensus around other areas of science and untrue statements of fact.

Analysis was undertaken to identify responses exceeding 86% consensus (i.e., that for global climate change among climatologists), not exceeding 5% non-consensus for refutation, and failing to refute the null hypothesis. Consensus margin of error calculations were done per the t-test by computing the margin of error for 86% and 5% consensus based on the number of respondents and size of the population with a Web-based calculator.[22] Similar calculations were done using the confidence interval for one proportion and sample size for one proportion, and they produced similar results.[23] It was identified that the scale applied (3-valued instead of a Likert scale) leads to an inability to validate the statistical characteristics using common methods.

No agreement reached 86% confidence levels or were within the margin of error (.77), and only a control question, #4 (∑a/N=.68), #5 (∑a/N=.75), and #9 (∑a/N=.64) (*N=54*) exceeded random levels of agreement. For disagreement, (*N=28*) only the same and one other control question, #5 (∑d/*N*=.14), and #9 (∑d/*N*=.10) were within the margin of error of not refuting consensus by disagreement (.05+.09=.14) levels. Only #1 (∑d/*N*=.53) and #11 (∑d/*N*=.50) were within random levels of refutation of consensus from disagreements. In summary, only #5 and #9 are viable candidates for overall community consensus of any sort, and those at levels of only 75% and 64% consensus respectively.[25]

The previous effort also involved a literature survey. 125 reviews of 95 unique published articles (31% redundant reviews) were undertaken. Of these, 34% are conference papers, 25% journal articles, 18% workshop papers, 8% book chapters, and 10% others. Publications surveyed included, without limit, IFIP (4), IEEE (16), ACM (6), HTCIA (3), Digital Investigation (30), doctoral dissertations (2), books, and other similar sources. A reasonable estimate is that there were less than 500 peer reviewed papers at that time that speak directly to the issues at hand. Results from examining 95 of those papers, which represent 19% of the total corpus, produces a 95% confidence level with a 9% margin of error. Of these reviews, 88% have no identified common language defined, 82% have no identified scientific concepts or basis identified, 76% have no identified testability criteria or testing identified, 75% have no identified validation identified, while 59% identify a methodology.

Internal consistency of these results was checked by testing redundant reviews to determine how often reviewers disagreed as to the “none” designation. Out of 20 redundant reviews (40 reviews, 2 each of 20 papers), inconsistencies were found for Science (3/20 = 15%), Physics (0/20 = 0%), Testability (4/20 = 20%), Validation (1/20 = 5%), and Language (1/20 = 5%). This indicates an aggregate error rate of 9/100 = 9% of entries in which reviewers disagreed about the absence of these indicators of scientific basis.

2 The present study

In order to differentiate between the problems associated with a lack of common terminology, language use, and methodological issues in the field, a short study was undertaken to try to differentiate actual consensus from linguistic issues. To do this, we created a survey that defines each term and measures agreement with the definition, and then tells the participant to assume the definition and evaluate the statement.

2.1 The survey methodology

As in the previous study, surveys were performed using the "SurveyMonkey" Web site. No identity-related data was collected or retained, although the survey mechanism prevents individuals from taking the survey from the same computer more than once unless they act to circumvent the mechanism. No accounting was taken to try to identify individuals who may have taken the survey as members of more than one group because the group overlap is relatively small. The new survey was introduced as follows:

> "This is a survey designed to identify, to a first approximation, whether or not there is a consensus in the scientific community with regard to some of the basic principles of the examination of digital forensic evidence.
>
> This survey is NOT about the physical realization of that evidence and NOT about the media in which it is stored, processed, or transported. It is ONLY about the bits.
>
> Please read carefully before answering.
>
> **For each numbered definition, review the definition and example(s) and indicate the extent to which you agree or disagree with the definition. For each follow-up, assume that the definition above it is correct, and respond to the statement in light of that definition, indicating the extent to which you agree or disagree with it.**

Following the instructions, the questions are provided in a format approximately as in Figure 1:

	Strongly disagree	***Disagree***	***Don't agree or disagree***	***Agree***	***Strongly Agree***
1: Definition: ...	⊙	⊙	⊙	⊙	⊙
Assuming the definition as the basis for your answer ...	⊙	⊙	⊙	⊙	⊙
2: Definition: ...	⊙	⊙	⊙	⊙	⊙

Figure 1 – The survey appearance

The set of questions and statements in the survey were as follows:

> *1: Definition: Digital forensics (as opposed to computer forensics) deals with sequences of bits and their use in legal actions (as opposed to attack detection or other similar areas). Example: A law suit or criminal charges with digital evidence will normally involve digital forensics. Do you agree with this definition?*
>
> *Assuming this definition as the basis for your answer, respond to the following statement: Digital evidence is only sequences of bits.*
>
> *2: Definition: Finite granularity means that, at the end of the day, you can only get so small, and no smaller. Finite granularity in space means that there is a minimum size (i.e., the bit) and finite granularity in time means that there is a minimum time unit (i.e., the number of bits that represent a time unit, or alternatively, the maximum clock speed of the mechanisms generating the bits). Example: At the level of digital evidence, as described earlier, no matter how much of it you have, there is no unit of that evidence smaller than a bit, and no matter has*

much precision measurements are made by, the times associated with those bits cannot be infinitesimally small. Do you agree with this definition?

Assuming this definition as the basis for your answer, respond to the following statement: Digital evidence is finite in granularity in both space and time.

3: Definition: Observation of digital information means to be able to determine what binary value each bit of the information has, while alteration is the changing of a bit value to the other bit value (digital information by definition only has two different values for bits). Example: At the level of digital evidence, when reading bits from a disk, the bits on the disk are observed by the mechanisms of the disk drive. When writing to a disk, some of the bits on the disk may be altered. Do you agree with this definition?

Assuming this definition as the basis for your answer, respond to the following statement: It is normally possible to observe digital information without altering it.

4: Definition: Duplication of digital information means to make an exact copy of the bit sequences comprising the digital information. Removal of digital information means to make alterations so that the original information (i.e., bit sequence) is no longer present where it originally was present before. Example: A duplicate of the bit sequence 1 0 1 will also be the bit sequence 1 0 1. The removal of the bit sequence 1 0 1 from a location would mean that the bit sequence at that location was no longer 1 0 1. Do you agree with this definition?

Assuming this definition as the basis for your answer, respond to the following statement: It is normally possible to duplicate digital information without removing it.

5: Definition: Trace evidence is evidence that is produced as the result of a process, so that the presence of the evidence is consistent with the execution of the process. Example: When a pen writes on paper, the indentations in the paper resulting from the writing are traces of the process of writing. Similarly, when a computer program produces bit sequences that are stored on a disk, the bit sequences are traces of the execution of the computer program that produced them. Do you agree with this definition?

Assuming this definition as the basis for your answer, respond to the following statement: Digital evidence is normally trace evidence.

6: Definition: Transfer is a concept in which two objects coming into contact with each other each leaves something of itself with the other. Example: When a shirt rubs up against a sharp object, some shards from the object may get transfered to the shirt while some fibers from the shirt may get transferred to the sharp object. Do you agree with this definition?

Assuming this definition as the basis for your answer, respond to the following statement: Digital evidence is normally not transfer evidence.

7: Definition: Latent evidence is evidence that cannot be directly observed by human senses. Example: DNA evidence is normally latent evidence in that the DNA has to be processed by a mechanism to produce a result that can be examined by human senses. Do you agree with this definition?

Assuming this definition as the basis for your answer, respond to the following statement: Digital evidence is normally latent in nature.

8: Definition: Computational complexity is the number of low-level computing operations required in order to perform an algorithm which processes information. Example: It takes more computer time to determine the best possible route from house to house throughout a city that to find any route that passes by all of those same houses. Do you agree with this definition?

Assuming this definition as the basis for your answer, respond to the following statement: Computational complexity limits digital forensic analysis.

9: Definition: A physics is a set of mathematical equations or other rules in context for describing and predicting behaviors of a system. Example: In the physical world, is is thought to be impossible to observe anything without altering it, because the act of observation alters the thing observed, and the physical world has no limits to granularity of space or time, so that no matter how small something is, there is always something smaller. Similarly, the physics of digital information, if you agree to statements above to that effect, is such that it is possible to observe bits without altering them, make an exact duplicate without altering the original, and so forth. Do you agree with this definition?

Assuming this definition as the basis for your answer, respond to the following statement: The physics of digital information is different than that of the physical world.

10: Definition: Consistency between two or more things means that each is the way you would expect it to be if the other ones are the way you observe them to be. Example: If you see a black box and someone else viewing the same object under the same conditions states that it is a white sphere, your observation is inconsistent with their statement. Similarly, if a sworn statement states that a particular file was created at 10AM on a particular day in a particular place, and the metadata for the file indicates that the same file was created at a different time on a different day, the sworn statement is inconsistent with the metadata. Do you agree with this definition?

Assuming this definition as the basis for your answer, respond to the following statement: As a fundamental of digital forensics, what is inconsistent is not true. (or in other words, the inconsistent things cannot all be true)

These correspond to questions 1, 3-11 of the previous survey, question 2 being removed because of the confusion surrounding it in discussions following the previous survey.

2.2 The raw data

The raw data from the survey is shown in Figure 2. Numbered (gray) columns correspond to definitions with the level of agreement to statements in the unnumbered (white) columns to their right. Rows represent responses, with responses 22-24 from the DFCB group and 1-21 from the IFIP group. At the bottom of the table there are two rows with responses from 2 other groups (Ignore 1 and Ignore 2), one response per group. The data from these groups was not included in the analysis because we could not adequately characterize these groups in terms of size or expertise, could not assure that they were independent of the DFCB and IFIP groups, and the number of samples is so small that no independent meaningful statistics can reasonably be gleaned. We will comment on their potential impact later.

In this table, 2 is “strongly disagree”, -1 is “disagree”, 0 is “don't agree or disagree”, 1 is “agree”, and 2 is “strongly agree”. In this table, a non-answer is treated as a “0”. At the end of the table (black background) are rows with calculations across columns (within responses or pairs of responses). LD is the level of disagreement between a definition and the relevant statement. That is, the number of instances where respondents agree (disagree) with the definition and disagree (agree) with the related statement (i.e., -1 or -2 for the definition and 1 or 2 for the statement or vice versa). A lower level of disagreement indicates a stronger correlation between the view of the definition and the view of the corresponding statement. “Agree” and “Disagree” are rows representing the number of responses agreeing (>0) and disagreeing (<0) respectively. Columns D, A, and DD represent the number of disagreements, agreements, and definition disagreements, respectively, for each respondent.

	1		2		3		4		5		6		7		8		9		10		D	A	DD
1	-1	1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-2	2	1	-2	-1	1	15	4	8
2	-1	1	-1	-1	1	1	-1	1	1	1	1	1	1	1	1	-1	1	1	-1	-1	7	13	4
3	-1	-2	0	0	-1	-2	-2	-1	1	1	-1	-2	0	0	-2	-2	-2	-2	-2	-2	14	2	7
4	-1	-1	0	-1	-1	-2	1	2	2	-2	2	-2	2	-2	2	1	-2	-2	2	-2	11	8	3
5	1	2	2	2	2	1	2	1	2	0	2	2	2	2	2	2	2	2	0	0	0	17	0
6	0	-2	0	1	1	-1	-2	-2	-1	-2	2	2	2	-2	2	0	1	-2	1	-1	9	8	2
7	1	-1	0	1	1	-1	-1	0	1	1	1	0	1	1	2	1	1	-1	1	-1	5	12	1
8	-2	-2	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	16	1
9	0	-1	-1	-1	2	2	-1	2	2	1	2	1	2	1	2	2	1	1	2	1	4	15	2
10	1	-1	1	1	1	1	1	1	2	2	-1	-1	1	1	1	1	1	1	1	1	3	17	1

	1		2		3		4		5		6		7		8		9		10		D	A	DD
11	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	20	0
12	1	0	1	1	-1	-1	-2	1	1	0	-1	1	-1	1	1	1	-1	1	1	0	6	11	5
13	2	2	2	2	2	2	2	2	2	2	1	2	2	2	1	2	1	2	2	2	0	20	0
14	2	-1	2	2	2	2	1	2	2	-1	2	-1	2	2	1	-1	-1	1	1	1	5	15	1
15	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	10	0
16	2	2	2	2	1	1	1	-1	2	2	2	1	2	2	2	2	2	2	2	1	1	19	0
17	0	0	1	0	1	2	1	2	1	1	0	1	1	0	0	0	1	0	1	0	0	11	0
18	-1	1	2	2	2	2	2	2	2	-1	2	-1	2	1	2	2	0	2	1	2	3	16	1
19	-2	-2	1	1	1	1	1	1	1	-1	1	1	1	1	1	0	1	1	1	1	3	16	1
20	-2	-2	0	0	1	1	1	0	1	-2	1	1	1	1	1	0	1	0	1	0	3	12	1
21	1	1	1	2	1	1	1	1	1	1	1	-1	1	1	1	2	1	2	1	1	1	19	0
22	2	1	1	1	1	0	0	1	1	1	1	0	0	1	0	1	1	-1	1	-1	2	13	0
23	-2	-2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	18	1
24	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	0	20	0
Agree	12	11	12	16	20	16	15	17	22	14	18	14	19	18	19	16	19	16	20	14		328	
Disagree	9	11	2	4	4	6	7	4	2	7	4	7	2	3	2	3	4	6	3	6	96		
LD		8		2		2		5		5		4		3		3		6		5			
Ignore 1	-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19	1
Ignore 2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	-1	1	1	1	1	1	19	0

Figure 2 - Data collected from the combined surveys

2.3 Analysis of survey results

This analysis covers the collection undertaken from the IFIP (*N=21*) and DFCB (*N=3*) groups, for a total population of 24 respondents. As depicted above, -2 is “strongly disagree”, -1 is “disagree”, 0 is “don't agree or disagree”, 1 is “agree”, and 2 is “strongly agree”. N is the number of respondents expressing an opinion (either agree or disagree), μ is the mean, and σ the standard deviation. Treating negative answers (-1, -2) as rejections of the asserted definition or statement (D=disagree), and positive answers (1 and 2) as affirmations of the asserted definition or statement (A=agree), we present the ratio of agreement (A/N) and disagreement (D/N) out of all respondents indicating a preference. The margin of error for 95% confidence for the identified sample sizes (from 17 to 24 out of a total estimated population of 250) is indicated under the column labeled M.[22] The C column indicates consensus above the margin of error for agreements (A) or disagreement (D) and is contains a “-” when no such consensus levels were found.

#	-2	-1	0	1	2	N	μ	σ	D	A	D/N	A/N	M	C	Issue
1	4	5	3	6	6	21	.21	1.44	9	12	.43	.57	.21	-	Definition
	6	5	2	6	5	22	-.04	1.51	11	11	.50	.50	.20	-	Only sequences of bits.
2	0	2	7	7	8	17	.88	.97	2	15	.12	.88	.23	A	Definition

#	-2	-1	0	1	2	N	μ	σ	D	A	D/N	A/N	M	C	Issue
	0	4	4	7	9	20	.88	1.09	4	16	.20	.80	.22	A	Finite granularity space and time.
3	0	4	0	12	8	24	1.0	1	4	20	.17	.83	.20	A	Definition
	2	4	2	8	8	22	.66	1.31	6	16	.27	.73	.20	A	Can observe bits w/out alteration.
4	3	4	2	9	6	22	.46	1.35	7	15	.32	.68	.20	-	Definition
	1	3	3	8	9	21	.88	1.17	4	17	.19	.81	.21	A	Can duplicate without removal.
5	0	2	0	11	11	24	1.29	.84	2	22	.08	.92	.20	A	Definition
	3	4	3	8	6	21	.42	1.35	7	14	.33	.67	.21	-	Digital evidence trace evidence.
6	0	4	2	8	10	22	1	1.08	4	18	.18	.82	.20	A	Definition
	2	5	3	8	6	21	.46	1.29	7	14	.33	.67	.21	-	Digital evidence not transfer.
7	0	2	3	8	11	21	1.17	.94	2	19	.10	.90	.21	A	Definition
	3	0	3	11	7	21	.79	1.22	3	18	.14	.86	.21	A	Digital evidence latent.
8	2	0	3	9	10	21	1.04	1.14	2	19	.10	.90	.21	A	Definition
	1	2	5	6	10	19	.92	1.15	3	16	.16	.84	.22	A	Computational complexity limits.
9	2	2	1	14	5	23	.75	1.13	4	19	.17	.83	.20	A	Definition
	4	2	2	8	8	22	.58	1.44	6	16	.27	.73	.20	A	Digital != real world physics.
10	1	2	1	14	7	24	.96	1.02	3	21	.14	.86	.20	A	Definition
	2	4	5	9	4	19	.38	1.18	6	13	.32	.68	.22	-	What is inconsistent is not true.

Figure 3 – Analysis of Consensus

Consensus above the margin of error from random is present for agreement with statements #2 (.80), #3 (.73), #4 (.81), #7 (.86), #8 (.84) and #9 (.73). This may be reasonably interpreted as indicating that for the full sample of the two organizations combined, there is a 95% chance that agreement would be above the margin of error from random for these 6 statements. In addition, agreement to #7 and #8 are within the margin of error of the 86% level of consensus seen among climatologists for global climate change.[18] That is, consensus was shown for:

Digital evidence is finite in granularity in both space and time.
It is normally possible to observe digital information without altering it.
It is normally possible to duplicate digital information without removing it.
Digital evidence is normally latent in nature.
Computational complexity limits digital forensic analysis.
The physics of digital information is different than that of the physical world.

And consensus was NOT shown for:

Digital evidence is only sequences of bits.
Digital evidence is normally trace evidence.
Digital evidence is normally not transfer evidence.
As a fundamental of digital forensics, what is inconsistent is not true. (or in other words, the inconsistent things cannot all be true)

A more important point is that consensus above the margin of error is present for 6 statements in the present study, whereas without the use of definitions in the survey, only two (corresponding to items

#4 and #8 in this survey) were above consensus.[25] In addition, consensus levels are higher (#4 went from .74 to .81 and #8 went from .64 to .84) when definitions were included.

It appears that a significant source of lack of consensus in the previous study was related to the lack of common language and agreed upon terminology in the field also identified in that study.[25]

Definitional disagreements are also worthy of commentary. While a few respondents (i.e., #1, #3, and #12) indicated disagreement or strong disagreement to at least half (5/10) of the definitions, only the definitions of what constitutes digital evidence (#1) and what constitutes duplication (#4) fail to reach consensus above random levels (8/10 definitions are at consensus levels above random). Definitions #2(.88), #5(.92), #7(.90), #8(.90) and #10(.86) (5/10) meet or exceed the 86% level of agreement for global climate change. There appears to be substantial disagreement regarding what constitutes digital evidence, and this is reflected in much of the methodological literature in the field. The lack of consensus levels for the definition of duplication is less clear based on the literature.

Analysis shows that of the 96 total disagreements, 49 of them (51%) stem from 4 respondents (#1, #3, #4, and #6). Without these respondents, consensus levels would be far higher. In addition, the two samples not included add only 2 disagreements, one of which is the definitional disagreement over what constitutes digital evidence, and the other a disagreement to statement 8. Adding these results in would drive consensus higher for all but statement 8 (which would go from 84% to 81%). None of these changes would result in moving non-consensus statements outside of the margin of error for randomness.

The internal level of disagreement (LD) between agreement on definitions and related statements is also of interest. Despite the lack of consensus around Definition #1, 8/24 respondents indicated different agreement to the definition than the statement. This suggests that respondents were able, as a group, to overcome differences in views on definitions to express views on statements in the context of the definitions provided. Among the 4 respondents constituting 51% of the disagreements, 2 of the 4 gave different answers to the statement than to the definition, again suggesting their ability to differentiate between their disagreement with the definitions and their agreement/disagreement with the statements. Looking at this more closely, only respondents #3, #8, and #23 always disagreed with statements when disagreeing with definitions. The skew of results toward “Agree” largely invalidates such an analysis regarding agreements, where 6 respondents agreed to all of the statements and definitions. Only respondent #11 had all identical answers. (strongly agree), indicating that respondents, as a whole, considered to some level their responses to each question.

3 Comments from reviewers

The review process yielded the following residual comments/questions which I address here. Reviewers seemed to comment on two basic issues; what the paper was about, and statistical issues.

The reviewers seemed to think that the paper was about the use of language and not a consensus around science. Comments included:

> *“the research sought people's opinions on generally accepted terminology in the field of digital forensics.”, “the [authors] claim that use of terminology is somehow a measure of science.”, “I think it is a bridge too far to suggest that there is no science (testability) without common terminology used by those in the discipline. ...”, “I think fundamentally that this paper has examined legal terminology, not underpinning science ...”*

This represents a misunderstanding of the purpose of the research and paper. Quoting from the abstract: “This paper updates previous work on the level of consensus in foundational elements of digital evidence examination. Significant consensus is found present only after definitions are made explicit, suggesting that, while there is a scientific agreement around some of the basic notions

identified, the use of a common language is lacking." The purpose was to mitigate the differences in use of language which were suspected as a partial cause of the lack of consensus identified in the previous paper. The question being addressed was whether the lack of identified consensus in prior research was due to an actual lack of agreement on the content or on differences in use of the language. This paper suggests that there is a lack of common definition that must be compensated for in order to measure consensus in this field, and that there is more consensus than previously thought.

The statistical question identifies correctly that 24 respondents is a seemingly small sample size. This is addressed by computing the margin of error for that sample size out of the total population, in this case estimated at 250. As such, this sample represents almost 10% of the total population and is proportionally a very large sample size compared to most statistical studies. Full details are provided so the reader can do further analysis and evaluate the actual responses using any desired method. The larger statistical problem is that the respondents are self-selected from the larger population, all of whom were notified of the study. We know of no way to compensate for this limitation through analysis and have no means to force compliance or expectation of gaining adequate samples from random polling.

4 Summary, Conclusions, and Further Work

It appears that this study confirms the hypothesis that the lack of consensus suggested in the previous study was due, at least in part, to a lack of common definitions and language in the digital forensics community. This study suggests that consensus is substantially present in many of the fundamental areas that are foundational to the acceptance of such evidence in legal proceedings. Clarity around definitions appears to be necessary for the field of digital forensics to reach levels of consensus present in other areas of science.

5 References

[1] R. Leigland and A. Krings, "A Formalization of Digital Forensics", International Journal of Digital Evidence, Fall 2004, Volume 3, Issue 2.
[2] Ryan Hankins, T Uehara, and J Liu, "A Comparative Study of Forensic Science and Computer Forensics", 2009 Third IEEE International Conference on Secure Software Integration and Reliability Improvement.
[3] Committee on Identifying the Needs of the Forensic Sciences Community, "Strengthening Forensic Science in the United States: A Path Forward", ISBN: 978-0-309-13130-8, 254 pages, (2009).; Committee on Applied and Theoretical Statistics, National Research Council.
[4] Scientific Working Group on Digital Evidence (SWGDE) Position on the National Research Council Report to Congress - Strengthening Forensic Science in the United States: A Path Forward
[5] S Garfinkel, P. Farrella, V Roussev, G Dinolt, "Bringing science to digital forensics with standardized forensic corpora", Digital Investigation 6 (2009) S2-S11
[6] M. Pollitt, "Applying Traditional Forensic Taxonomy to Digital Forensics", Advances in Digital Forensics IV, IFIP TC11.9 Conference Proceedings, 2009.
[7] G. Carlton and R. Worthley, "An evaluation of agreement and conflict among computer forensics experts", Proceedings of the 42nd Hawaii International Conference on System Sciences, 2009
[8] NIST, "Computer Forensics Tool Testing (CFTT) Project", http://www.cftt.nist.gov/
[9] The Federal Rules of Evidence, Section 702.
[10] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 125 L. Ed. 2d 469, 113 S. Ct. 2786 (1993).
[11] Frye v. United States, 293 F 1013 D.C. Cir, 1923
[12] Reference Manual on Scientific Evidence - Second Edition - Federal Judicial Center, available at http://air.fjc.gov/public/fjcweb.nsf/pages/16

[13] U.S. Department of Justice, “A Review of the FBI’s Handling of the Brandon Mayfield Case”, unclassified executive summary, January 2006. (http://www.justice.gov/oig/special/s0601/exec.pdf)
[14] K. Popper, The Logic of Scientific Discovery (1959), Hutchins and Company, London. ISBN10: 0415278449.
[15] J. Jones and D. Hunter, “Qualitative Research: Consensus methods for medical and health services research”, Volume 311, Number 7001, BMJ 1995; 311 : 376 (5 August 1995).
[16] Karin D. Knorr, “The Nature of Scientific Consensus and the Case of the Social Sciences”, in Karin D. Knorr, Karin Knorr-Cetina, Hermann Strässer, Hans-Georg Zilian, “Determinants and controls of scientific development”, Institut für Höhere Studien und Wissenschaftliche Forschung (Vienna, Austria), pp 227-256, 1975.
[17] A. Fink, J. Kosecoff, M. Chassin, and R. Brook, “Consensus Methods: Characteristics and Guidelines for Use”, AJPH September 1984, Vol. 74, No. 9.
[18] Margaret R. K. Zimmerman, “The Consensus on the Consensus: An Opinion Survey of Earth Scientists on Global Climate Change”, Dissertation, 2008.
[19] North Eastern Forensics Exchange, Georgetown University, 8/13 – 8/14, 2010.
[20] Forensics Data Base is available at http://calsci.org/ under the “FDB” menu selection.
[21] Edmond Locard and D. J. Larson, “The Analysis of Dust Traces” (in 3 parts), The American Journal of Police Science, V1 #4, 1930.
[22] A calculator from http://www.raosoft.com/samplesize.html was used to perform this calculation, based on the Z value method, which is imprecise at sample sizes under 30, but close enough for the purposes applied.
[23] Lenth, R. V. (2006-9). Java Applets for Power and Sample Size [Computer software]. Retrieved 2010-09-27 from http://www.stat.uiowa.edu/~rlenth/Power.
[24] Cole, Simon A. “Out of the Daubert Fire and Into the Frying Pan? Self-validation, meta-expertise, and the admissibility of Latent Print Evidence in Frye Jurisdictions”, Minn. Journal of Law, Science, and Technology, V9#2, pp 453-541, 2008.
[25] Bar-Anan, Yoav; Wilson, Timothy D.; Hassin, Ran R., “Inaccurate self-knowledge formation as a result of automatic behavior.”, J. of Experimental Social Psychology, V46, #6, pp 884-895, 2010
[26] F. Cohen, J. Lowrie, and C. Preston, “The State of the Science of Digital Evidence Examination”, IFIP TC11, Jan 2011.
[27] F. Cohen, “Putting the Science in Digital Forensics“, Journal of Digital Forensics, Security and Law, Vol. 6(1) 7 Column 1, July, 2011.