Chapter 1

THE STATE OF THE SCIENCE OF DIGITAL EVIDENCE EXAMINATION

Fred Cohen, Julie Lowrie and Charles Preston

Abstract This paper examines the state of the science and the level of consensus in the digital forensics community regarding digital evidence examination. The results of this study indicate that elements of science and consensus are lacking in some areas and are present in others. However, the study is small and of limited scientific value. Much more work is required to evaluate the state of the science of digital evidence examination.

Keywords: Digital evidence examination, science, status

1. Introduction

There have been increasing calls for scientific approaches and formal methods in digital forensics (see, e.g., [7, 8, 11, 16, 17, 19]). At least one study [3] has shown that, in the relatively mature area of evidence collection, there is a lack of agreement among and between the technical and legal communities about what constitutes proper process. The National Institute of Standards and Technology [15] has tested various tools used in digital forensics, including tools for evidence collection. The results show that the tools have substantial limitations about which digital forensics professionals must be aware if reliable results are to be assured.

Meanwhile, few, if any, efforts have focused on understanding the state of the science in digital evidence examination (i.e., analysis, interpretation, attribution, reconstruction and aspects of presentation). This paper describes the results of preliminary studies of the state of scientific consensus in the digital forensics community regarding digital evidence examination in the context of the legal mandates.

2. Legal Mandates

The U.S. Federal Rules of Evidence (FRE) [22] and the rulings in the Daubert [23] and Frye [20] cases express the most commonly applied standards with respect to expert witnesses. 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, only expert witnesses can address issues based on scientific, technical and 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 (i) the testimony is based on sufficient facts or data; (ii) the testimony is the product of reliable principles and methods; and (iii) 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 upon cross-examination.

The Daubert standard [23] essentially allows the use of accepted methods of analysis that reliably and accurately reflect the data on which they rely. The Frye standard [20] focuses on: (i) whether the findings presented are generally accepted within the relevant field; and (ii) 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 U.S. and much of the world are based on principles and specific rulings that dictate, in essence, that the evidence be: (i) beyond the normal knowledge of non-experts; (ii) based on a scientific methodology that is testable; (iii) characterized in specific terms with regard to reliability and rates of error; (iv) processed by tools that are properly tested and calibrated; and (v) consistent with a scientific methodology that is properly applied by the expert as demonstrated by the information provided by the expert [5, 20, 22, 23].

Failure to meet these requirements can be spectacular. In the Madrid bombing case, the U.S. FBI declared that a fingerprint from the scene demonstrated the presence of an Oregon attorney. However, this individual, after having been arrested, was clearly demonstrated to have been on the other side of the world at the time in question [21]. The side-effect is that fingerprints are now challenged as scientific evidence around the world [4].

3. 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 the circumstances where the refutation applies). A statement about a universal principle can be disproved by a single refutation, but any number of confirmations cannot prove it to be universally true [18].

In order to make scientific statements regarding digital evidence, there are some deeper requirements that may have to be met. In particular, there has to be some underlying common language that allows scientists to communicate 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 assess test results). These ultimately have 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 the requirements are met. Consensus may be tested by surveying individuals who are active in a field (e.g., individuals who testify as expert witnesses and publish in relevant peer-reviewed venues) regarding their understandings to see whether and to what extent there is a consensus in the field. Polling has been used in a number of fields to assess consensus [6, 9, 10]. For example, a recent survey [24] seeking to measure consensus in the field of Earth science noted that more than 86% of Earth scientists agreed with and less than 5% disagreed with the claim that human activity is a significant contributing factor to global climate change.

4. Preliminary Studies

In order to understand the state of the science, we performed two limited studies, both of them preliminary in nature. These studies were not undertaken with a high level of scientific rigor, the intent being to suggest the state of the science of digital evidence examination, not to definitively demonstrate it.

4.1 Informal Poll

A very limited and informal poll was conducted at an NSF/ACM sponsored workshop on digital forensics (Northeastern Forensics Exchange, Georgetown University, Washington, DC, August 2010) to expose the audience to issues related to scientific consensus in the field, and to obtain a preliminary assessment of the level of agreement among individuals who self-assert that they are undertaking scientific research or actively working in the field. The attendees included academics who actively teach or conduct research in digital forensics, and funding agency representatives, government researchers and industry professionals who specialize in digital forensics. A total of 31 individuals were present during the polling. Fifteen of them self-identified themselves as scientists who perform research in the field, and five indicated that they had testified in a legal matter as a digital forensic expert.

All the attendees who identified that they had taken a physics course indicated that they had heard of the equation $F = ma$, and that they agreed, in most cases, that this equation was reliable for the identified purpose (100%). Note that a failure to agree does not indicate disagreement. This demonstrates a consensus among attendees that they: (i) had heard of this physics principle and (ii) agree to its validity in the appropriate circumstances.

Five attendees indicated that they had heard of the Second Law of Thermodynamics. Four of them agreed to its validity in the appropriate circumstances (80%). Again, this represents some level of scientific consensus.

When asked if the speed of light limited how fast physical objects could travel in the normal universe, eighteen of the twenty attendees (90%) who had heard of the concept agreed with it. Again, this represents some level of consensus in an area most physicists would consider basic knowledge.

Two "made up" physics principles were introduced as control questions. Only one individual indicated he/she had heard about one of these principles.

The attendees were notified that the issues to be discussed dealt only with digital evidence, not physical evidence. Therefore, the focus would be on bits and not the media that contain, transport or process them or the underlying physical characteristics of the media. For each concept, the attendees were polled on whether they had previously heard of the concept (H) and, of those, how many agreed with it (A). Table 1 summarizes the poll results.

Table 1. NSF/ACM poll results.

#	**Concept**	**H**	**A**	**%**
1	Digital evidence is only sequences of bits	7	7	100
2	The physics of digital information is different than that of the physical world	5	1	20
3	Digital evidence is finite in granularity in both space and time	6	4	66
4	Observation of digital information without alteration	12	9	75
5	Duplication of digital information without removal	12	9	75
6	Digital evidence is trace evidence	14	5	35
7	Digital evidence is not transfer evidence	0	0	–
8	Digital evidence is latent in nature	2	1	50
9	Computational complexity limits digital forensic analysis	12	12	100
10	Theories of digital evidence examination form a physics	2	1	50
11	The fundamental theorem of digital forensics is "What is inconsistent is not true"	3	2	66

To the extent that this unscientific polling of workshop attendees may be of interest, it suggests that, while there is a level of scientific consensus ($\geq$80%) among attendees claiming to have limited knowledge of physics about some of the basic concepts of physics, a similar level of consensus does not exist for a similar set of basic principles in digital forensics. Interestingly, only four out of the eleven concepts had previously been heard of by more than half of the self-asserted scientists and experts who responded ($n = 14$). Of the four concepts, only one concept is at a consensus level similar to the attendees' consensus about physics ($\geq$80%). Widely-recognized concepts that are central to the admissibility of evidence and that have been widely accepted by the courts, (i.e., Concepts #4 and #5) are agreed upon by only 75% of the attendees who had heard of them. The basic notion that digital evidence is trace evidence is agreed upon by 35% of the attendees who had heard of the concept. These results do not (and could not) indicate a consensus similar to that for the physics concepts, because a failure to agree cannot be interpreted as disagreement. In this sense, the poll was asymmetric.

By way of comparison, refutation of the null hypothesis in psychology generally requires a 95% level of certainty, while the global climate change consensus mentioned above was accepted at the 86% level. The only consensus in the group of polled attendees was that computational

complexity limits digital forensic analysis. Thus, while the poll is hardly a valid scientific study of the issues, it suggests that the null hypothesis (i.e., there is no scientific consensus regarding digital forensics) is confirmed.

4.2 Online Surveys

The results of the initial poll demonstrated the need for further study. A survey methodology was applied in which the same or very similar statements in similar order were presented to different populations from the digital forensics community. Members of the Digital Forensics Certification Board (DFCB), members of the International Federation of Information Processing (IFIP) Working Group 11.9 on Digital Forensics, and members of the Bay Area Chapter of the High Tech Crime Investigators Association (HTCIA) were solicited for participation in the surveys.

The DFCB consists of 165 certified practitioners, all of whom have substantial experience in digital forensics, including more than five years of professional experience and experience testifying as experts in legal proceedings. A total of 80 DFCB members were solicited for the survey.

The IFIP Working Group 11.9 members come from around the world. They include academics, active duty law enforcement personnel, corporate computer crime investigators, researchers and others. Most, if not all, have published peer-reviewed papers in digital forensics, and many have testified as expert witnesses in legal matters. Some overlap exists between the IFIP and DFCB groups.

The HTCIA membership consists of peace officers, investigators and attorneys engaged in the investigation or prosecution of criminal activities associated with computer systems and networks, and senior corporate security professionals. The Bay Area HTCIA Chapter has about 80 members who are active in digital forensics. Few, if any, of the Bay Area HTCIA members are DFCB practitioners, and none are IFIP members. Thus, the three groups, while not strictly mutually exclusive, are substantially independent in terms of membership.

Survey participation was solicited via email. Each survey appeared on a single web page with one item per line. The DCFB online survey instructions are shown in Figure 1. Each line in the DFCB survey had a checkbox on the same line for "I've heard of it" and "I agree with it."

The instructions for the HTCIA and IFIP surveys are shown in Figure 2. The instructions are slightly different from those for the DFCB survey to accommodate the fact that each statement had three checkboxes for

Forensic Science Consensus – 2010

This is a simple survey designed to identify, to a first approximation, whether or not there is a consensus in the scientific community with regard to 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.

- Don't look anything up. Only go from what you already know.

- If you haven't heard of the principle/concept, don't agree with it!

- These are not necessarily all true or false. Only go with what you know.

- This is ONLY about digital evidence – not its physical realization.

- Agreement means that it is normally the case when dealing with digital evidence, not a universal truth.

- EXCEPTIONS: Items marked (Physics) are about the normal physics of time and space.

Figure 1. DFCB online survey instructions.

"I disagree," "I don't know" and "I agree," from which one choice had to be made.

The three surveys used the SurveyMonkey website; each survey was up for five days. No identity-related data was collected or retained. However, the survey mechanism prevents respondents from taking the survey from the same computer more than once. Attempts were not made to identify respondents who may have taken the survey as members of more than one group; this is because group overlaps are very small, if at all.

Table 2 lists the survey statements. Note that the first column (#) was not included in the actual survey. Statement #A is a well-known physics equation; any individual who has had a high school physics course has likely encountered and applied this equation. Statement #B is a control question, designed to detect if boxes are checked automatically (e.g., by computer programs), without reading or disingenuously; there is no such equation in physics. If random guessing were used, there would be a 75% chance of triggering one or the other or both of the responses

Forensic Science Consensus – 2010

This is a simple survey designed to identify, to a first approximation, whether or not there is a consensus in the scientific community with regard to 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.

- Don't look anything up. Only go from what you already know.

- These are not necessarily all true or false. Only go with what you know.

- This is ONLY about digital evidence – not its physical realization.

- "I agree" means it is normally the case when dealing with digital evidence, not a universal truth.

- "I disagree" means it is normally not the case when dealing with digital evidence, not that it can never be true.

- "I don't know" means you haven't heard of it or don't agree or disagree with it.

- EXCEPTIONS: Items marked (Physics) are about the normal physics of time and space.

Figure 2. IFIP and HTCIA online survey instructions.

to Statement #B, and, thus, most random guesses would be detected. Statement #C is widely agreed upon by the physics community, but not as well-known in the general community; it is assumed not to be true in many science fiction works. All three physics questions would likely receive universal agreement among physicists: Statement #A would be heard of and agreed to, Statement #B would not be heard of or agreed to, and Statement #C would be heard of and agreed to.

Statements #C and #9 are also related in that Statement #C may "prime" [1] Statement #9. Similarly, Statement #3 has the potential to prime Statements #4, #5, #6 and #9. Also, because the survey allows changes, Statements #4, #5, #6 and #9 have the potential to prime Statements #3 and #10. Finally, Statements #3 and #10 should be internally consistent within respondents.

Table 2. Statements used in the online surveys.

#	**Statement**
A	$F = ma$ (Physics)
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
B	The Johnston-Markus equation dictates motion around fluctuating gravity fields (Physics)
6	Digital evidence is trace evidence
7	Digital evidence is not transfer evidence
8	Digital evidence is latent in nature
C	Matter cannot be accelerated past the speed of light (Physics)
9	Computational complexity limits digital forensic analysis
10	Theories of digital evidence examination form a physics
11	The fundamental theorem of digital forensics is "What is inconsistent is not true"

Note that the statements in Table 2 have the same labels as the equivalent statements in the poll (Table 1). The nature of the NSF/ACM poll and the DFCB online survey is that results do not and cannot indicate a consensus against these concepts, because a failure to agree cannot be interpreted as disagreement. In this sense, the survey statements are asymmetric, just like the poll statements. Note also that the IFIP and HTCIA online surveys fail to differentiate "I don't know" from "I never heard of it."

Table 3 shows the results of the original poll and the three subsequent surveys, along with the summary results. The highlighted rows labeled #A, #B and #C correspond to the control statements. The study groups are in columns (from left to right): shaded for the NSF/ACM (N) poll (n = 14), unshaded for the DFCB (D) survey (n = 11), shaded for the IFIP (I) survey (n = 23), unshaded for the HTCIA (H) survey (n = 2) and shaded for the summaries ($\sum$). For N and D, the columns are "I've heard of it" (H), "I agree with it" (A), percentage agreeing (% = 100*A/H) and A/n. For I and H, the columns are "I disagree" (d),

S#	NH	NA	%	A/n	DH	DA	%	A/n	Id	Ia	%	d/n	a/n	Hd	Ha	%	d/n	a/n	∑a	∑d	a/N	d/N
A	22	22	100	n/a	8	6	75	.50	2	17	89	.08	.73	0	0	0	0	0	37	2	.68	.07
1	7	7	100	.50	9	6	66	.50	13	10	76	.56	.43	2	0	0	1.0	0	23	15	.42	.53
2	5	1	20	.07	3	2	66	.17	9	12	57	.39	.52	0	1	50	0	.50	16	9	.29	.32
3	6	4	66	.28	2	1	50	.08	6	16	72	.26	**.69**	1	1	50	.50	.50	22	7	.40	.25
4	12	9	75	**.64**	10	10	100	**.83**	6	17	73	.26	**.73**	1	1	50	.50	.50	37	7	.68	.25
5	12	9	75	**.64**	12	11	92	**.92**	3	20	86	.13	**.86**	1	1	50	.50	.50	41	4	.75	.14
B	1	0	0	0	0	0	0	0	1^+	2^+	0^+	0^+	0^+	0	0	0	0	0	na	na	na	na
6	14	5	35	.35	8	4	50	.33	6	14	70	.26	.60	1	1	50	.50	.50	24	7	.44	.25
7	0	0	0	0	5	2	40	.17	5	6	54	.21	.26	1	1	50	.50	.50	9	6	.16	.21
8	2	1	50	.07	5	3	60	.25	5	13	72	.21	.56	1	1	50	.50	.50	18	6	.33	.21
C	20	18	90	n/a	10	4	40	.33	2	14	87	.08	.60	1	0	0	.50	0	32	3	.59	.10
9	12	12	100	**.85**	4	3	75	.24	3	18	85	.13	**.78**	0	2	100	0	1.0	35	3	.64	.10
10	2	1	50	.07	1	1	100	.08	9	7	43	.39	.30	1	0	0	.50	0	9	10	.16	.35
11	3	2	66	.14	0	0	0	0	13	7	35	.43	.30	1	1	50	.50	.50	10	14	.18	.50

Figure 3. Results of polling and the online surveys.

"I agree" (a), percentage of decided agreeing (% = 100*a/(a+d)), a/n and d/n.

In the case of the IFIP and HTCIA surveys, the Control Statement #B is 66.7% likely to detect problems if answered ("d" or "a" are problems). The analysis of the results in Table 3 demonstrates consensus views and within the margin of error for not refuting consensus views of different survey groups and of the survey as a whole using the consensus level for global climate change (e.g., total population of around 5,000, n = 1,749, p = .88, margin of error = 1.9% for 95% certainty) [24]. This appears to be adequate to establish scientific consensus, regardless of the controversy surrounding the particulars of the study. Thus, $\geq$.86 of the validated sample will be considered to represent a "consensus."

4.3 Analysis of Results

It appears that about half of the DFCB survey respondents chose either "H" or "A" instead of "H" or "H and A." As a result, responses identifying only "A" are treated as having received "H and A." This issue is addressed in the subsequent IFIP and HTCIA surveys by allowing only "I agree," "I disagree" and "I don't know."

An analysis was undertaken to identify the responses exceeding 86% consensus, not exceeding 5% non-consensus for refutation, and failing to refute the null hypothesis. Consensus margin of error calculations were performed as a t-test by computing the margin of error for 86% and

5% consensus based on the number of respondents and the size of the population.

Similar calculations were performed using the confidence interval for one proportion and the sample size for one proportion; the calculations produced similar results. The margin of error calculations are somewhat problematic because: (i) the surveys have self-selected respondents and are, therefore, not random samples; (ii) normality was not and cannot be established for the responses; and (iii) a margin of error calculation assumes the general linear model, which is not validated for this use. The margin of error is valid for deviations from random guesses in this context and, thus, for confirming the null hypothesis with regard to consensus, again subject to self-selection.

The NSF/ACM poll had a maximum of fourteen respondents (n = 14) for non-physics questions. Assuming that there are 50 comparable individuals in the U.S., the margin of error is 23% for a 95% confidence level. Given a level of agreement comparable to that supporting global climate change (A/n $\geq$ .86) [24], only Statement #9 (100%, A/n = .85) is close. Statements #4 and #5 (75%, A/n = .64) are barely within the margin of error ([.41, .87] $\geq$ .86) of not refuting consensus at 95% confidence and refuting consensus at 90% confidence (margin of error = .19). Only Statement #9 (A/n = .85) is differentiable from random responses beyond the margin of error (.50 + .23 = .73).

The DFCB online survey had twelve respondents (n = 12). For a population of 125 and an 86% A/n consensus level, a 95% confidence level has a margin of error of 28%. The DFCB survey responses demonstrate that, while there are high percentages of agreement among respondents who have heard of Statements #4 (100%, A/n = .83) and #5 (92%, A/n = .92), only Statement #5 meets the consensus level of global climate change while Statement #4 is within the margin of error. Control Statement #B properly shows no responses, and there is no overall agreement on Control Statement #A (75%, A/n = .50). Only Statements #4 (A/n = .83) and #5 (A/n = .92) are differentiable from random responses beyond the margin of error (.50 + .28 = .78).

The IFIP survey had 26 respondents, three of whom were eliminated because of "a or d" responses to Statement #B (n = 23). For a population of 128 and an 86% a/n consensus level, a 95% confidence level has a margin of error of 19%. The IFIP survey responses demonstrate consensus for Statement #5 (86%, a/n = .86, d/n = .13) and response levels within the margin of error for Statements #3 (72%, a/n = .69, d/n = .26), #4 (73%, a/n = .73, d/n = .26) and #9 (85%, a/n = .78, d/n = .13). None of the denied response counts are below the refutation consensus level (d/n $\leq$ .05) of the global climate change study [24],

which tends to refute consensus. The best refutation consensus levels are for Control Statements #A and #C (d/n = .08). Statements #3 and #4 have refutation rates (d/n = .26) beyond the margin of error for consensus (.26 - .19 > .05). Thus, of the statements within the margin of error but not at the consensus level, only Statement #9 remains a reasonable candidate for consensus at the level of the global climate change study. Only the responses to Statements #3 (a/n = .69), #4 (a/n = .73), #5 (a/n = .86) and #9 (a/n = .78) have acceptance that is differentiable from random beyond the margin of error (.50 + .19 = .69). Failure to reject beyond the margin of error (.50 – .19 = .31) is present for Statements #A (d/n = .08), #3 (d/n = .26), #4 (d/n = .26), #5 (d/n = .13), #6 (d/n = .26), #7 (d/n = .21), #8 (d/n = .21), #C (d/n = .08) and #9 (d/n = .13). Therefore, these statements are not refuted from possible consensus at the 95% level by rejections alone, and only Statements #3, #4 and #9 are viable candidates for consensus beyond random levels.

The HTCIA survey had only two respondents (n = 2). The margin of error for this sample size is approximately 75%, so the responses are meaningless for assessing the level of consensus.

Combining the online survey results yields the summary columns in Table 3. Because there are two different question sets, combining them involves different total counts. For A and a (agreement numbers), the total number of respondents is 54 (N = 54) and the total population is 382, yielding about a 9% margin of error for an 86% confidence level. For d (disagreement numbers), the total count is 28 ($N = 28$) and the total population is 208, yielding a margin of error of 13% for an 86% confidence level. No agreement reaches the 86% confidence level or is within the margin of error (.77), and only Statements #A ($\sum$a/N = .68), #4 ($\sum$a/N = .68), #5 ($\sum$a/N = .75) and #9 ($\sum$a/N = .64) exceed random levels of agreement. For disagreement, only Statements #A ($\sum$d/N = .07), #5 ($\sum$d/N = .14), #C ($\sum$d/N = .10) and #9 ($\sum$d/N = .10) are within the margin of error of not refuting consensus by disagreement levels (.05 + .09 = .14). Only Statements #1 ($\sum$d/N = .53) and #11 ($\sum$d/N = .50) are within random levels of refutation of consensus from disagreements. In summary, only Statements #5 and #9 are viable candidates for overall community consensus of any sort, with consensus levels of only 75% and 64%, respectively.

4.4 Literature Review for Scientific Content

The second study (which is ongoing) involves a review of the published literature in digital forensics for evidence of the underlying elements

of a science. In particular, we are reviewing the literature in digital forensics to identify the presence or absence of the elements of science identified above (i.e., that a common language for communication is defined, that scientific concepts are defined, that scientific methodologies are defined by or used, that scientific testability measures are defined by or scientific tests are described, and that validation methods are defined by or applied).

To date, we have undertaken 125 reviews of 95 unique publications (31% redundant reviews). Of these, 34% are conference papers, 25% journal articles, 18% workshop papers, 8% book chapters and 10% others. The publications include IFIP (4), IEEE (16), ACM (6) and HTCIA (3) publications, *Digital Investigation Journal* articles (30), doctoral dissertations (2), books and other similar publications. A reasonable estimate is that there are less than 500 peer-reviewed papers today that speak directly to the issues at hand. Results from examining 95 of these papers, which represent 19% of the total corpus, produces a 95% confidence level with a 9% margin of error.

Of the publications that were reviewed, 88% have no identified common language defined, 82% have no identified scientific concepts or basis identified, 76% have no testability criteria or testing identified and 75% have no validation identified. However, 59% of the publications do, in fact, identify a methodology.

The results were checked for internal consistency by testing redundant reviews to determine how often reviewers disagreed with the "none" designation. Out of the twenty redundant reviews (40 reviews, two each for twenty 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% ($= 9/100$) of entries in which reviewers disagreed about the absence of these scientific basis indicators.

Primary and secondary classifications of the publications were generated to identify, based on the structure defined in [2], how they might best be described as fitting into the overall view of digital forensics and its place in the legal system. Primary classifications (one per publication) for this corpus were identified as 26% legal methodology, 20% evidence analysis, 8% tool methodology, 8% evidence interpretation, 7% evidence collection, and 31% other (each less than 4%). Secondary classifications (which include the primary classification as one of the identifiers and are expressed as the percentage of reviews containing the classification, so that the total exceeds 100%) were identified as 28% evidence analysis, 20% legal methodology, 19% tool methodology, 15% evidence collection, 12% evidence interpretation, 10% tool reliability, 10% evi-

dence preservation, 9% tool testing, 9% tool calibration, 9% application of a defined methodology, and 7% or less of the remaining categories.

The internal consistency of category results was tested by comparing major primary areas for redundant reviews. Of the twenty redundant reviews, two have identical primary areas and sub-areas (e.g., Evidence:Preserve), four have identical areas but not sub-areas (e.g., People:Knowledge and People:Training) and the remaining thirteen have different primary areas (e.g., Challenges:Content and Evidence:Interpret). For this reason, relatively little utility can be gained from the exact categories. However, in examining the categories from redundant reviews, no glaring inconsistencies were identified for the chosen categories (e.g., Evidence:Analyze with Process:Disposition).

Full details of these reviews, including paper titles, authors, summaries and other related information are available at [2]. The corpus and the reviews will expand over time as the effort continues.

A reasonable estimate based on the number of articles reviewed and the relevant publications identified is that there are only about 500 peer-reviewed science or engineering publications in digital forensics. While a sample of 95 is not very large, it constitutes about 20% of the entire digital forensics corpus and the results may be significant in this light. While the classification process is entirely subjective and clearly imperfect, the results suggest an immature field in which definitions of terms are not uniformly accepted or even well-defined. Issues such as testability, validation and scientific principles are not as widely addressed as in other areas. Also, there appears to be a heavy focus on methodologies, which may be a result of a skewing of the source documents considered, but it seems to suggest that digital forensics has not yet come to a consensus opinion with regard to methodologies. Many researchers may be defining their own methodologies as starting points as they move toward more scientific approaches.

Longitudinal analysis has not yet been performed on the available data, and it is anticipated that such an analysis may be undertaken once the data is more complete. Early indications based on visual inspection of the time sequence of primary classifications suggest that methodology was an early issue up to about 2001 when evidence analysis, interpretation, and attribution became focal points, until about 2005, when methodology again became a focus, until the middle of 2009, when analysis started to again become more dominant. These results are based on a limited non-random sample and no controls for other variables have been applied. They may, as a matter of speculation, be related to external factors such as the release of government publications, legal rulings

or other similar things in the field of forensics in general or in digital forensics as an emerging specialty area.

4.5 Peer Reviews

Three peer reviews of this paper provided qualitative data worthy of inclusion and discussion. The reviewers primarily commented on the survey methodology, questions and the statistical analysis.

Comments on the survey methodology were of two types, technical and non-technical. The technical comments have been addressed in this paper. The non-technical comments surrounded the use of the physics questions and their selection. The physics questions were used as controls, a common approach when no baselines exist.

Comments on the survey questions covered three issues. First, the questions do not represent areas where there is a consensus. Second, knowing the correct answers to the questions does not necessarily mean that digital forensic tasks are performed properly. Third, the questions are unclear and they use terminology that is not widely accepted.

Statistical comments focused on the utility of the comparison with global climate change and the validity of statistical methods in this context. The validity issues are discussed in the body of this paper, but whether or not there is utility in comparing the results with consensus studies in other fields is a philosophy of science issue. This study takes the position that a level of consensus that is above random is inadequate to describe the state of a science relative to its utility in a legal setting. The only recent and relevant study that we found was on global climate change. This is an issue of which the public and, presumably, jury pools, attorneys and judges would be aware. Thus, it is considered ideal for this study dealing with the legal context.

The presence or absence of consensus was the subject of the study, so the assertion that the questions represent areas where there is a lack of consensus is essentially stating that the results of the study reflected the reviewers' sense of the situation. This is a qualitative confirmation of the present results, but begs the question of whether there are areas of consensus. A previous study [3] has been conducted on this issue for evidence acquisition and consensus was deemed to be lacking. However, the issue was not examined in the same manner as in the present study.

The question of whether and to what extent understanding the underlying physics and mechanisms of digital forensics is required to perform forensic examinations and testify about them is interesting. At the NSF/ACM sponsored workshop where our poll was conducted, the NSF representative indicated that the NSF view was that digital forensics is

a science like archeology and not like physics. This begs the question of whether archeologists might need to understand the physics underlying carbon dating in order to testify about its use in a legal setting. This paper does not assume that the survey questions are important *per se*, but the lack of consensus for questions such as whether evidence can be examined without alteration or without the use of tools suggests that these issues are likely to be challenged in legal settings [20, 22, 23].

The assertion that the terminology is unclear or not widely accepted in the field is, in fact, the subject of the study, and the peer reviews again confirm the null hypothesis regarding consensus. In essence, digital forensic practitioners do not even agree on what the questions should be considered to determine whether there is a consensus regarding the fundamentals of the field.

As qualitative data points, the peer reviews appear to confirm the results of the paper. The fact that this paper was accepted after peer reviews suggests that the reviewers recognize the consensus issue as important and problematic at this time.

5. Conclusions

The two preliminary studies described in this paper individually suggest that: (i) scientific consensus in the area of digital forensic evidence examination is lacking in the broad sense, but that different groups in the community may have limited consensus in areas where they have special expertise; and (ii) the current peer-reviewed publication process is not helping bring about the elements typically found in the advancement of a science toward such a consensus. Publication results also suggest that methodologies are the primary focus of attention and that, perhaps, the most significant challenge is developing a common language to describe the field. This is confirmed by the substantial portion of "I don't know" responses in the consensus surveys. The peer reviews of a earlier version of this paper also qualitatively support these results.

Our studies are ongoing and the results may change with increased completeness. The surveys to date have small to moderate sample sizes and the respondents are self-selected from the populations they are supposed to reflect. Also, the highly interpretive and qualitative nature of the paper classification approach is potentially limiting.

The margins of error in the surveys are 19% to 27%. The surveys involved approximately 10% of the total populations of authors of peer-reviewed articles, 10% of the certified digital forensics practitioners in the United States, 10% of the professors teaching digital forensics at the graduate level in U.S. universities, and a smaller percentage of investiga-

tors in the field. Another measure is the control statements, which had better consensus levels among the participants who are not, as a rule, self-asserted experts, performing scientific research or publishing peer-reviewed articles in physics. This suggests that the level of consensus surrounding digital evidence examination is less than that surrounding the basics of physics by non-physicists. While this is not surprising given the relative maturity of physics, it appears to confirm the null hypothesis about scientific consensus around the core scientific issues in digital evidence examination. Yet another measure is the levels of refutation shown in the IFIP and HTCIA surveys. Not only was consensus largely lacking, but substantially higher portions of the populations expressed that the asserted principles were not generally true and refuted them. The only candidates for overall community consensus beyond the random level and not refuted by excessive disagreements are Statement #5 (75% consensus) "It is possible to duplicate digital information without removing it" and Statement #9 (64% consensus) "Computational complexity limits digital forensic analysis." These levels of consensus appear to be lower than desired for admissibility in legal proceedings.

Some of the survey results are disconcerting given that there have been many attempts to define terms in the field, and there is a long history of the use of some of the terms. For example, the notions of trace, transfer and latent evidence have been used in forensics since Locard almost 100 years ago [12–14]; yet, there is a lack of consensus around the use of these terms in the survey. This suggests a lack of historical knowledge and thoroughness in the digital forensics community.

Future work includes completing the preliminary review of the literature and performing more comprehensive studies of scientific consensus over a broader range of issues. Also, we intend to undertake longitudinal studies to measure progress related to the building of consensus over time. As an example, once the literature review is completed, results over a period of several years could be analyzed to see if changes over this period have moved toward an increased use of the fundamental elements of science identified in this paper.

References

[1] Y. Bar-Anan, T. Wilson and R. Hassin, Inaccurate self-knowledge formation as a result of automatic behavior, *Journal of Experimental Social Psychology*, vol. 46(6), pp. 884–895, 2010.

[2] California Sciences Institute, Forensics Database (FDB), Livermore, California (calsci.org).

[3] G. Carlton and R. Worthley, An evaluation of agreement and conflict among computer forensics experts, *Proceedings of the Forty-Second Hawaii International Conference on System Sciences*, 2009.

[4] S. Cole, Out of the Daubert fire and into the Fryeing pan? Self-validation, meta-expertise and the admissibility of latent print evidence in Frye jurisdictions, *Minnesota Journal of Law, Science and Technology*, vol. 9(2), pp. 453–541, 2008.

[5] Federal Judicial Center, Reference Manual on Scientific Evidence (Second Edition), Washington, DC (www.fjc.gov/public/pdf.nsf/lookup/sciman00.pdf/$file/sciman00.pdf), 2000.

[6] A. Fink, J. Kosecoff, M. Chassin and R. Brook, Consensus methods: Characteristics and guidelines for use, *American Journal of Public Health*, vol. 74(9), pp. 979–983, 1984.

[7] S Garfinkel, P. Farrell, V. Roussev and G Dinolt, Bringing science to digital forensics with standardized forensic corpora, *Digital Investigation*, vol. 6(S), pp. 2–11, 2009.

[8] R. Hankins, T. Uehara and J. Liu, A comparative study of forensic science and computer forensics, *Proceedings of the Third IEEE International Conference on Secure Software Integration and Reliability Improvement*, pp. 230–239, 2009.

[9] J. Jones and D. Hunter, Qualitative research: Consensus methods for medical and health services research, *British Medical Journal*, vol. 311(7001), pp. 311–376, 1995.

[10] K. Knorr, The nature of scientific consensus and the case of the social sciences, *International Journal of Sociology*, vol. 8(1/2), pp. 113–145, 1978.

[11] R. Leigland and A. Krings, A formalization of digital forensics, *International Journal of Digital Evidence*, vol. 3(2), 2004.

[12] E. Locard, The analysis of dust traces – Part I, *American Journal of Police Science*, vol. 1(3), pp. 276–298, 1930.

[13] E. Locard, The analysis of dust traces – Part II, *American Journal of Police Science*, vol. 1(4), pp. 401–418, 1930.

[14] E. Locard, The analysis of dust traces – Part III, *American Journal of Police Science*, vol. 1(5), pp. 496–514, 1930.

[15] National Institute of Standards and Technology, Computer Forensics Tool Testing Program, Gaithersburg, Maryland (www.cftt.nist.gov).

[16] National Research Council of the National Academies, *Strengthening Forensic Science in the United States: A Path Forward*, National Academies Press, Washington, DC, 2009.

[17] M. Pollitt, Applying traditional forensic taxonomy to digital forensics, in *Advances in Digital Forensics IV*, I. Ray and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 17–26, 2008.

[18] K. Popper, *The Logic of Scientific Discovery*, Hutchins, London, United Kingdom, 1959.

[19] 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, Document 2009-09-17 (www.swgde.org/documents/current-documents), 2009.

[20] U.S. Circuit Court of Appeals (DC Circuit), Frye v. United States, *Federal Reporter*, vol. 293, pp. 1013–1014, 1923.

[21] U.S. Department of Justice, A Review of the FBI's Handling of the Brandon Mayfield Case, Office of the Inspector General, Washington, DC (www.justice.gov/oig/special/s0601/exec.pdf), 2006.

[22] U.S. Government, Federal rules of evidence, Title 28 – Judiciary and Judicial Procedure Appendix and Supplements, *United States Code*, 2006.

[23] U.S. Supreme Court, Daubert v. Merrell Dow Pharmaceuticals, Inc., *United States Reports*, vol. 509, pp. 579–601, 1983.

[24] M. Zimmerman, The Consensus on the Consensus: An Opinion Survey of Earth Scientists on Global Climate Change, M.S. Thesis, Department of Earth and Environmental Sciences, University of Illinois at Chicago, Chicago, Illinois, 2008.