The State of the Science of Digital Evidence Examination

Fred Cohen, Julie Lowrie, and Charles Preston

dr.cohen@all.net julie@lowrie.net charles.preston@mac.com

California Sciences Institute

572 Leona Drive, Livermore, CA 94550

Abstract—*This paper suggests the state of the science and level of consensus in digital evidence examination. Elements of science and consensus are found lacking in some areas and present in others, but the studies involved are of only limited scientific value, and much further work is needed.*

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]

The same sort of effort has not been found to date in understanding the state of the science in digital evidence examination (i.e., analysis, interpretation, attribution, reconstruction, and aspects of presentation). This paper describes preliminary results of limited studies of the state of scientific consensus in digital evidence examination in the context of the legal mandates of the US Federal Rules of Evidence [9] and relevant case law.

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. 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. The latter 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]

2 Preliminary studies performed and results

In order to understand the state of the science, we performed two limited studies, both preliminary, one ongoing, the other of small sample size, and neither undertaken with a high level of scientific rigor. These studies are intended to suggest the state of the science, not to definitively demonstrate it.

2.1 A limited poll at a workshop

A very limited poll (attendees raising hands) was taken of attendees at a National Science Foundation (NSF) Association for Computing Machinery (ACM) sponsored workshop on digital forensics [19] for the purposes of both enlightening the audience as to the underlying scientific consensus issues of the field, and to get a preliminary gauge on the level of agreement between people who self-assert that they are undertaking scientific research or actively working in the field. Speakers included 17 professors actively teaching or researching in digital forensics, funding agency representatives, government researchers in the field, and industry workers specializing in digital forensics. At the time

of the poll, 31 attendees in total were present, 15 self-identified as scientists performing research in the field, and 5 indicated that they had testified in a legal matter as an expert in digital forensics.

All attendees who identified that they had taken a physics course indicated that they had previously heard of the equation “F=ma” and that they agreed that, in most cases, this equation was reliable for identified purposes. (100%) Note that a failure to agree does not indicate disagreement. This demonstrates a consensus among attendees that they (1) had heard of this physics principle and (2) agree to its validity in appropriate circumstances. Of 5 self-identifying that they had heard of the 2nd law of thermodynamics, 4 agreed to its validity in appropriate circumstances. (80%) Again, this represents some level of scientific consensus. A third question asking whether the speed of light limited how fast physical objects could go in the normal universe produced 18 of 20 (90%) who had heard of the concept and 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 and only one individual indicated they had ever heard of one of them.

Attendees were notified that the issues to be discussed dealt only with digital evidence and not physical evidence, and that therefore, we were discussing only bits and not the media that contain, transport, or process them or the underlying physical characteristics of that media. For each concept, participants were polled as to how many had previously heard of the concept (H) and, of those, how many agreed with it (A). Results are summarized in Table 1:*

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	n/a
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

Table 1 – Poll results

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 (≥80%) among attendees claiming limited knowledge of physics, about some of the basic concepts of physics, a similar level of consensus in the field where they claim expertise does not exist across a similar set of basic principles. Regardless of whether one view or another of the actual truth of the concepts identified may end up as part of a scientific consensus around digital forensics examination, it is clear that today, only 4 of 11 concepts had previously been heard of by more than half of the self-asserted scientists and experts in the field (n=14). Of those 4, only one concept was at a consensus level similar to their consensus for physics (≥80%). Widely recognized concepts that are central to the admissibility of forensic evidence and that have been widely accepted by the courts, (i.e., 4 and 5) were only 75% agreed upon by those who had heard of them. The basic notion that digital evidence is trace evidence was agreed upon by 35%

of those who had heard of the concept. These results also do not and could not indicate a consensus similar to that of physics against these concepts, because a failure to agree cannot be interpreted as disagreement. In that sense, the poll was asymmetric.

By way of comparison, refutation of the null hypothesis in psychology generally requires a 95% level of certainty, and global climate change is accepted at the 86% level. The only consensus in this group was that computational complexity limits what analyses can be done. Thus, while this 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.

2.2 An extension of the initial poll in other communities

Based on preliminary results from the initial poll, further study seemed warranted. A survey methodology was applied in which the same or very similar statements in similar order were asked in surveys to different populations from the digital forensics community. 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, and members of a Bay Area chapter of the High Tech Crime Investigators Association (HTCIA).

The DFCB consists of 165 certified practitioners, all of whom have substantial experience in digital forensics, including being accepted as testifying experts in legal matters and professional work in the field for more than 5 years. 80 DFCB members received the solicitation for this survey. The IFIP digital forensics conference attendees come from all over the world. They include academics, active duty law enforcement personnel, corporate computer-related 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. There is an overlap of a few individuals who are both IFIP conference attendees and DFCB certified practitioners. The HTCIA consists of Peace officers, investigators and prosecuting Attorneys engaged in the investigation and/or prosecution of criminal activity associated with computers and/or advanced technologies and senior security professionals who control and are responsible for security and/or investigation in computer or advanced technology environments. The Bay Area chapter has about 80 members and is quite active in digital forensics, it appears that few if any of them are DFCB certified practitioners, and that none of them attended IFIP conferences from which the IFIP list was generated. Thus the groups, while not strictly mutually exclusive, are substantially independent in terms of membership.

Invitations were in the form of an email stating:

> I am taking a very brief survey (anonymous in every way I can easily make it) to get a sense of the presence or absence of a consensus around basic issues in digital forensics.
>
> If you would be interested in taking this survey, please go to the URL below, check off the boxes that apply, and press the completion button.
>
> The whole survey should take only a minute or two to complete, and I don't want you to look anything up. Just answer with what you already know or don't.
>
> THE URL WENT HERE
>
> Thanks for your assistance.
> FC

Surveys appeared on a single Web page consisting or one item per line. For the DFCB survey, each with a checkbox on the same line for each of “I've heard of it” and “I agree with it.”. The instructions on the Web site were:

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.

For the HTCIA and IFIP surveys 3 checkboxes per statement, only one of which could be selected, offered “I disagree.”, “I don't know.”, and “I agree.” The instructions deleted the line starting with “ If you haven't heard” and replaced the line starting with “Agreement means” with the following 3 lines:

- “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.

The specific statements are identified in Table 2 (column **L** were not included in the actual survey):

L	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 forensic evidence form a physics.
11	The fundamental theorem of digital forensics is "What is inconsistent is not true".

Table 2 – Statements used in the online polls

Surveys were performed using the “SurveyMonkey” Web site for 5 days per survey. 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.

Statement #A is a well known definition from physics. Anybody who has had high school physics has likely encountered and applied this equation. Statement #B was made up as a control question to detect instances where boxes were checked automatically (e.g., by computer programs), without reading, or disingenuously. There is no such equation in physics today. If random guessing were used, there would be a 75% chance of triggering one or the other or both of the responses to #B, and thus a significant quantity of survey results from such mechanisms would likely be detected. Statement #C is widely agreed upon by the physics community but not as well known in the general community, and assumed not to be true in many works of science fiction. All of the physics questions would likely receive universal agreement among professional physicists (#A would be heard of and agreed to, #B would not be heard of or agreed to, and #C would be heard of and agreed to).

Statements #C and #9 are also related in that #C may prime [25] #9. Similarly, #3 has the potential to prime #4, #5, #6, and #9, and because the survey allows changes, #4, #5, #6, and #9 have the potential to prime #3 and #10. #3 and #10 should be internally consistent within respondents.

The statements labeled with numbers indicate the same numbering as for equivalent statements in Table 1 and the ordering and placement of the control statements are similar to the ordering used in the original poll. The nature of the NSF/ACM and DFCB surveys is that results do not and can not indicate a consensus against these concepts, because a failure to agree cannot be interpreted as disagreement. In this sense, these survey statements, just as the poll questions, were asymmetric. The IFIP and HTCIA surveys fail to differentiate “don't know” from “never heard of it”.

Table 3 shows the results of the original poll, the three subsequent surveys, and summary results. Highlighted rows labeled A, B, and C are the control statements. The study groups are in columns left to right, shaded for NSF/ACM (N, n=14), unshaded for DFCB (D, n=11), shaded for IFIP (I, n=23), unshaded for HTCIA (H, n=2), and shaded for summaries (∑). For N and D, columns include “I've heard of it” (H), “I agree with it” (A), percentage (%) agreeing (100*A/H), and A/n. For I and H, columns include “I disagree.” (d), “I agree” (a), percentage (%) of decided agreeing100*a/(a+d), a/n, and d/n. Summary details are described below. For the IFIP and HTCIA surveys, control statement #B is 2/3 likely to detect problems if answered (“d” or “a” are problems). Analysis following 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 ~5000, n=1749, p=.88, margin of error=1.9% for 95% certainty) [18] that appears to be adequate to establish scientific consensus, regardless of the controversy surrounding the particulars of that study. Thus ≥.86 (of the validated sample) will be considered “consensus”.

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

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*
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

Table 3 – Results of the poll and 3 surveys*

2.3 Analysis of results

It appears that about half of the survey respondents for the DFCB chose either “H” or “A “rather than “H” or “H and A”. As a result, responses identifying only “A” are treated as having received “H and A”. This was addressed for HTCIA and IFIP by allowing only Agree, Disagree, and “I don't know.”.

Analysis was undertaken to identify responses exceeding 86% consensus, 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] The margin of error calculations are somewhat problematic in this application because (1) the surveys have self-selected respondents and are thus not random samples, (2) normality has not been and cannot be established for responses, and (3) 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 14 respondents (n=14) for non-physics questions. Assuming there are 50 comparable individuals in the US, for a 95% confidence level, the margin of error is 23%. Given a level of agreement comparable to that supporting global climate change (A/n≥.86) [18], only #9 (100%, A/n=.85) is close, while #4 and #5 (75%, A/n=.64), are barely within the margin of error ([.41 - .87]≥.86) of not refuting consensus at 95% confidence, and refute consensus at 90% confidence (margin of error=.19). Only #9 (A/n=.85) was differentiable from random responses beyond the margin of error (.50+.23=.73).

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

The IFIP survey had 26 respondents, 3 of which were removed because of a or d responses to #B (n=23). For a population of 128 and an 86% a/n consensus level, a 95% confidence level result gives a margin of error of 19%. [22] IFIP responses demonstrate consensus level for #5 (86%, a/n=.86, d/n=.13) and response levels within the margin of error for #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 were below the refutation consensus level (d/n≤.05) of global climate change[18], which tends to refute consensus. The best refutation consensus levels were for controls #A and #C (d/n=.08), and #3 and #4 had refutation rates (d/n=.26) beyond the margin of error for consensus (.26-.19>.05). Thus, of those within the margin of error but not at consensus levels, only #9 remains a reasonable candidate for consensus at the level of global climate change. Items #3 (a/n=.69), #4 (a/n=.73), #5 (a/n=.86), and #9 (a/n=.78) were the only answers with acceptance differentiable from random beyond the margin of error (.50+.19=.69). Failure to reject beyond the margin of error (.50-.19=.31) was present for #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), so these are not refuted from possible consensus at the 95% level by rejections alone, and #3, #4, and #9 remain viable candidates for consensus beyond random levels.

The HTCIA survey had 2 respondents (n=2). At this sample size, the margin of error is approximately 75%, and thus these numbers are meaningless in terms of consensus.

Combining survey results produced the summary columns of Table 3. Because the two survey types have different question sets, combining them uses different total counts. For A and a (agreement numbers) the total number of respondents was 54 (N=54) and the total population size was 382, producing about 9% margin or error for an 86% confidence level. For d (disagreement numbers) the total count was 28 (*N*=28) and the total population size was 208, producing a margin of error of 13% for an 86% confidence level. No agreement reached 86% confidence levels or were within the margin of error (.77), and only #A (∑a/N=.68), #4 (∑a/N=.68), #5 (∑a/N=.75), and #9 (∑a/N=.64) exceeded random levels of agreement. For disagreement, only #A (∑d/*N*=.07), #5 (∑d/*N*=.14), #C (∑d/*N*=.10), 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.

2.4 A literature review for scientific content

The longer of the two efforts involves an ongoing review of the published literature in the field for evidence of the underlying elements of a science. In particular, the authors of this study undertook and are continuing to undertake a review of literature in digital forensics with a specific focus on identifying 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 in the publication, that scientific testability measures are defined by or scientific tests described by the publication, and that validation methods are defined by or applied within the publication).

To date, this effort has undertaken 125 reviews of 95 unique published articles (31% redundant reviews). 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 are less than 500 peer reviewed papers today 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, but 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.

Primary and secondary classifications of the articles were generated to identify, based on the structure defined in [20], how they might best be described as fitting into the overall view of digital forensics and its place within the legal system. Primary classifications (1 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 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% evidence preservation, 9% tool testing, 9% tool calibration, 9% application of a defined methodology, and 7% or less of the remaining categories.

Internal consistency of category results was tested by comparing major primary areas for redundant reviews. Out of 20 redundant reviews, 2 had identical primary areas and sub-areas (e.g., Evidence:Preserve), 4 had identical areas but not sub-areas (e.g., People:Knowledge and People:Training), and the remaining 13 had different primary areas (e.g., Challenges:Content and Evidence:Interpret). For this reason relatively little utility regarding the exact categories can be gained. However, in examining the categories from redundant reviews, no glaring inconsistencies between the categories chosen (e.g., Evidence: Analyze with Process: Disposition) were identified.

Full details of these reviews, including the summary of the papers, authors, titles, and other related information is available and will continue to expand over time as the effort continues. [20]

Based on these preliminary numbers, some analysis is worth considering. 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 the field. While a sample of 95 is not very large, it constitutes about 20% of the entire corpus in this area, and the results may be significant in that context. 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, where issues such as testability, validation, and scientific principles have not been as widely addressed as might be found in other areas, where underlying scientific methodologies are not being regularly and rigorously applied, and where there is a lack of consensus surrounding even the basic issues. There appears to be a heavy focus on identifying methodologies, which may be a result of a skewing of the source documents considered, but it seems to suggest that the field has not yet come to a consensus opinion with regard to methodologies. Many researchers are defining their own methodologies in order to find a place to start to move toward a more scientific approach.

Longitudinal analysis has not yet been performed on the available data, and it is anticipated that such 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.

2.5 Peer Reviews

In addition to technical changes reflected throughout the body of the paper, the peer reviews (n=3) provided qualitative data worthy of inclusion and discussion. Roughly, the reviews indicated three sorts of information; (1) comments on methodology, (2) comments on the questions, and (3) comments on the statistics, all largely focused on the surveys.

Statistical comments surrounded the utility of comparison to global climate change and the validity of statistical methods in this context. The validity issues are identified in the body of the paper, but whether there is utility in comparing results to consensus studies of other fields is a philosophy of science issue. This study takes the position that levels of consensus above random are inadequate to describe the state of a science relative to its utility in legal settings. The only recent study we found on point was on global climate change, and this is an issue that the public, and presumably jury pools, lawyers, and judges, are aware of. As such it is considered ideal for understanding in legal contexts.

Comments on methodology were of two types. Technical comments were integrated into the paper. Non-technical comments surrounded the use of the “physics” questions and the selection of the questions. The physics questions were used as controls, a common approach when no previous baselines exist. The selection of questions as a methodological question is actually the same as the comments on the questions outside of the asserted realm of methodology.

Comments on questions surrounded three issues, to wit; (1) the questions don't represent areas where there is a consensus, (2) knowing the correct answers are not necessary to doing digital forensics properly, and (3) the questions were unclear and used terminology that was not widely accepted.

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. Previous study has been done on this issue for acquisition[7] and consensus was lacking, but not examined in the same way as in the present study.

The question of whether and to what extent knowing of and understanding the underlying physics and mechanisms of digital forensics is required in order to do forensic examinations and testify about them is an interesting one. At the same conference at which the first survey was done, the NSF representative indicated that the NSF view was that digital forensics is a science like archeology and not like physics. This then 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. The present paper does not assume that the survey questions were “important” per se, but the inability to gain consensus around questions such as whether evidence can be examined without alteration or without the use of tools suggests issues likely to be challenged in legal settings. [9][10][11]

The assertion that the terminology was unclear or not widely accepted in the field was in fact the subject of the study, and as such, the peer reviews again confirm the null hypothesis regarding consensus. In essence, as a field, digital forensics practitioners don't even agree on what the questions should be to determine whether there is a consensus around the basics of the field.

The peer reviews, as qualitative data points, appear to confirm the results of the paper. The fact that the paper was accepted with these peer review comments, along with other comments in those reviews, suggests that the reviewers recognize the consensus issue as important and problematic at this time.

3 Summary, Conclusions, and Further Work

These two preliminary studies individually suggest that (1) scientific consensus in the area of digital forensic evidence examination is lacking in the broad sense, but that different groups within that overall community may have limited consensus around areas in which they have special expertise, and (2) that the current peer-reviewed publication process is not acting to bring about the sorts of elements typically found in the advancement of a science toward such a consensus. Publication results also suggest that methodologies are a substantial focus of attention and that perhaps the most significant challenge may be in the development of a common language to describe the field. This is confirmed by the substantial portion of “don't know” responses to consensus surveys. The peer reviews also qualitatively support these results.

Study is ongoing and results may change with increased completeness. The methodologies applied for getting these results have small to moderate sample sizes, respondents are self-selected from the populations they are supposed to reflect, and the highly interpretive nature of the paper classification approach and qualitative nature of that study are potentially limiting. The surveys produced margins of error of 19-27%. The surveys in total involved something like 10% of the total populations of peer reviewed literature authors, 10% of the certified digital forensics practitioners in the US, 10% of the professors teaching graduate programs in this area in the US, and a smaller percentage of the members of investigators 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 per non-physicists. While this should not be a surprise given the relative maturity of physics compared to digital evidence examination, it seems to confirm the null hypothesis about scientific consensus around the core scientific issues in digital evidence examination. Still 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 population expressed that the asserted principles were not generally true and refuted them. The only candidates for overall community consensus at beyond random levels of agreement and not refuted by excessive disagreements were #5 (75% consensus) “It is possible to duplicate digital information without removing it.” and #9 (64% consensus) “Computational complexity limits digital forensic analysis.” These levels of consensus appear to be lower than desired for admissibility in legal proceedings, and they are far from the level of consensus of other science that remains highly contested in the general public.

Some of the survey results are particularly disconcerting given that there have been various 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 and their particulars have been widely recognized and used in forensics since the time of Locard almost 100 years ago,[21] yet there was 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 that is not representative of sound research and demonstrates a lack of general acceptance in the field.

Further work includes completing the preliminary review of literature and performing more comprehensive studies of scientific consensus over a broader range of issues. It seems clear that additional work toward consensus among and between these groups might include (a) texts that accurately reflect the historical terms and uses in the field and increased requirements of rigor in the use of terminology in peer reviewed publications, (b) the definition and widespread promulgation of a common language for the field reflected in the literature and review processes for publication, (c) the creation of a better set of reference sources and educational process to move toward consensus, and (d) the creation of a common body of knowledge that is backed up by theory and experiments that are

widely repeated and taught universally. These efforts might substantially help to bring scientific consensus that would allow the field to become more widely accepted and uniformly applied.

Finally, it would seem sensible to undertake longitudinal studies to measure progress of the building of consensus over time. As an example, once the literature review is completed, results over a period of several years may be analyzable to see if changes over that period have moved toward an increased use of the fundamental elements of science identified above. Similarly, more comprehensive and repeated studies of certified examiners, conference attendees, peer reviewed publication editors and reviewers, and similar groups, might be prudent.

* The percentage and other figures throughout this paper are of higher precision at 2 digits than the accuracy, which never exceeds 1/denominator in the ratios as presented).

+ Survey responses containing these particular answers were removed from the data set and are not represented in the sample size or other calculations.

4 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