

WHO WILL TEST THE TESTERS?

David Harley, Andrew Lee

ESET Research, 610 West Ash Street, Suite 1900,
San Diego, CA 92101, USA

Tel +1 619 876 5458

Email {dharley, alee}@eset.com

ABSTRACT

The anti-malware industry has been plagued since its earliest days by one poorly designed comparative test after another. In 2007, some of the best anti-malware researchers, comparative testers and product certification specialists took the first steps towards raising product testing standards with the formation of a group specifically focused on establishing standards and methodologies, educating both consumers and testers in discrimination between good and bad practice, and providing objective analyses of current testing practices. This paper summarizes current initiatives by the Anti-Malware Testing Standards Organization and other groups, but also considers next steps, going beyond objectifying methodology, educational issues and blowing away the fog of misinformation and fallacy, to the next level. Underlying these vital issues is a question: is it possible to make testers and certifying authorities more accountable for the quality of their testing methods and the accuracy of the conclusions they draw based on that testing?

This paper attempts to answer that question.

INTRODUCTION

There is no doubt in the minds of those who have the challenging task of designing, creating and maintaining anti-malware solutions, that their products are extraordinarily complex beasts whose manufacture requires a high degree of professional knowledge, hard work and technical expertise.

Contrast this then with the rather surprising fact that anyone with a computer, Internet access, a (perhaps rather undefined) 'collection' of malware, and a little interest in the subject can open up shop as a computer anti-virus tester, and thereby claim to be a champion of the user and nemesis of the bloated and increasingly beleaguered anti-malware product vendors.

Such *chutzpah* would be met, in any other field, by not a little disdain, and perhaps a few lawsuits. Imagine an untrained lone practitioner setting up shop as a 'tester' of antibiotics, artificial hearts or perhaps more appropriately, prophylactics.

However, in the field of computer security there seems to be a rather pervasive opinion that anti-malware products can be tested in much the same way as any other piece of software, or even hardware, that any result thus generated is valid, and any complaints arising from the industry as a result of such tests are counted as sour grapes from bad losers who are too lazy and stupid to make decent products. So our happy new tester merrily gathers together a collection of 'samples', gets his digital fingers on a few anti-malware products, runs a few scans and publishes the results – blissfully unaware that in most cases, those 'results'

are, at best, misleading, and at worst are downright wrong, and damaging both to the consumer and to the vendors tested. This sort of testing is certainly a part of the reason that the anti-malware industry in general suffers a poor public image [1].

Endorsement (of tests or testers) by the anti-malware industry is often popularly discounted as irrelevant and even counter-productive, and indeed people who sell an anti-malware product *do* have a vested interest in exercising influence over how their products are tested. However, most testers cannot test detection performance properly without the co-operation of the industry, a primary source of validated samples. It may not be necessary to be vendor-aligned or a specialist researcher to test detection effectively – though independent testing may be more useful to the end-user – but detection testing with large sample sets is more difficult without industry trust. Of course, software testing *in general* is difficult, and it causes us all some concern that there is a widespread belief that it's any easier to test anti-malware products than other kinds of software and system, requiring no special knowledge of testing or of anti-malware technologies.

There have been many papers and presentations [2–4], discussions and 'late night in the bar' debates about anti-malware testing – frequently between aggrieved malware researchers bemoaning their mistreatment at the hands of some misguided tester – but only since early 2007 has there been cohesive industry action towards establishing a standards body dedicated to improving testing practice. This culminated in the formation of AMTISO (Anti Malware Testing Standards Organization) [5], with a remit to create independent, vendor- and technology-neutral standards, guidelines and best practices for the testing community. It is noteworthy that this body includes representation of many of the more established testing organizations, who have themselves recognized the damage sustained by their own reputations from poor testing performed by others without the required specialist knowledge.

(BAD) PRACTICE MAKES IMPERFECT

Good and bad practice in testing security products has been of major concern to anti-malware researchers since the early days of comparative testing and product certification. In 2000, many researchers signed a letter [6] in protest against a particularly poor test, and there have been other responses (some piecemeal, some coordinated) to other misconceived tests and commentary. These include the *Consumer Reports* test of 2006 [7], the *Untangled* test of 2007 [8], and misleading commentary from other sectors of the security industry [10, 11]. These responses presented and clarified the anti-malware industry's view of specific tests, but had little impact on the media or public, security amateurs and specialists in other security fields, or subsequent testers falling into similar traps. Indeed, they often meet with hostile reactions, less from testers under fire than from vociferous instant experts [12, 13].

However, 2007 saw a more concerted effort to standardize and improve methodologies used by established testing organizations and to mitigate the impact of poorly designed and frankly misleading tests from other sources. In May 2007, a major anti-virus testing workshop assembled vendors and product

testers to talk about ways of improving test methods [4]. Over time, anti-malware researcher responses to poor tests were semi-spontaneously coordinated across vendor and researcher blogs and widely discussed [10, 14, 15]. Several papers at the AVAR 2007 conference dealt with testing issues [3, 16, 17], and testers and vendors were worked together towards improved testing, addressing the need for behavioural analysis methodologies in security software [5].

Here, we summarize recent organizational developments in addressing these issues, but also consider human and informational aspects of raising standards by:

- Providing better information on how (not) to test, and how to evaluate tests by others.
- Making testers and certifying authorities more accountable for the quality of their methodology and the accuracy of their conclusions.
- Countering misinformation/misinterpretation propagated by third-party commentators inside and outside the security industry.

The range of known viruses detected by mainstream anti-virus scanners is fairly consistent, but the range of other detections (trojans, rootkits, bots, spyware, and so on) and other functionalities differs widely, particularly in terms of newly discovered and non-viral threats. Yet there is a continuing need for enterprises and individuals to achieve a balance between affordability and effectiveness in detection and removal of many of these threats, and multiple test types (comparative review testing, product certification testing) are considered significant aids to assessment and evaluation.

Unfortunately many (e.g. journalists, consumer groups and security amateurs) have very publicly used inappropriate methodologies for testing detection-focused software, to the annoyance of several parties:

- The vendors whose products are disadvantaged by ill-founded test criteria.
- Vendors who fare better, but are conscious that being well-placed in a poor test can actually damage credibility.
- Testers who have invested years in developing expertise and establishing the resources necessary for exhaustive and well-founded testing.
- Anyone wanting testing to benefit consumers, rather than to mislead them [10].

Vendors acknowledge that sound testing provides invaluable feedback on how well a given product functions in the real world. Poor testing, on the other hand, only favours testers with a hidden agenda and vested interests [3], or as a means of boosting magazine sales with spectacular stories of passes and failures.

Outside the industry, there is a tendency:

- To underestimate the difficulties of sound testing
- To overestimate the value of informal, biased, incompetent testing
- To overestimate the average person's ability to evaluate a test

- To underestimate the competence of AV and overestimate the degree to which an AV researcher's judgement is impaired by marketing and self-protection
- To fail to understand the ethical considerations that should drive a tester.

CATALYSTS

Flash points in the past year or so have included:

- Lingering resentment over the 2006 *Consumer Reports* test [7] and the 'AntiVirus FightClub' test at LinuxWorld by *Untangle* [18].
- A report [19] concerning a behaviour-analysis-focused test by a popular German magazine.

The *Consumer Reports* test is seen in the industry as having been based on an unsound testing strategy [20]. Perhaps, though, the endorsement of the test as 'fair and rigorous' by SANS [9] did even more damage, despite vigorous attempts at remediation [11].

SANS was not the first (or last) to misrepresent modern anti-malware software as being purely reactive, but its commentary also made damaging and unprovable assumptions about the validity of the test methodology. Access to the results was restricted, and the tests were outsourced to a third party who declined to comment because of the risk of 'involvement' [20]. Unfortunately, industry commentators were perceived as reacting (as usual) to the 'ethically' inappropriate behaviour of creating test viruses [21, 22]. It might be useful for the industry to focus more on the technical objections [10] to test methodology, where actually ascertainable, given that many people outside the industry don't understand or sympathize with the ethical objections [21, 22]. However, it's a cause for concern when an organization perceived as authoritative in the field of information security uncritically accepts third-hand information as a basis for critical commentary.

The *Untangle* 'Antivirus FightClub' did at least make some attempt to make its methodology open and reproducible, albeit in ways that only encouraged the unwary to make the same errors [10]. This was actually a three-part test:

- Testing with the EICAR test file, which the tester persisted in calling a virus. Unfortunately, useful though this tool sometimes is, it's not the first time it has been misused in testing [24].
- The second test used an 'in-the-wild' test set 'which we picked from my mailbox ... [8]'. The use of the term 'in the wild' in this case applied to samples of unproven validity (the tester never answered questions as to how his samples were validated) was one of the drivers for the launching of the WildList Organization Working Group (discussed below).
- The third phase used a test set supplied by members of the audience at the test venue, 'which ranged from pretty standard viruses to some bizarre stuff I couldn't identify' [8].

The evident shortcomings of this test have been explored at greater length elsewhere [10].

The magazine review of December 2007 [19] was cited in a rant about the inadequacy of behaviour analysis in current anti-malware products, and also referenced an earlier defence of the *Consumer Reports* test [22], criticizing the industry's ethical objections to virus creation for testing purposes. The details of the test remain unavailable, so we cannot assess the technical competence of the testing. The articles imply that the tests used a virus generator and/or manual code modification. These approaches present technical objections, implying that the presentation of the samples does not represent a real-life scenario. More recently, the 'Race to Zero' competition scheduled for DEFCON 16, though not overtly described as a comparative test (though it purports to show, once again, the general uselessness of current anti-malware solutions and will make comparisons between the products used for the test), raised similar issues and extreme reactions much discussed at, among other places, <http://www.eset.com/threat-center/blog/>.

THE REACTION

In May 2007, a CARO Antivirus Testing Workshop assembled vendors and comparative testers to talk about ways of improving test methods. The presentations available at <http://www.f-prot.com/workshop2007/presentations.html> demonstrate the commitment to improvement at this workshop, with presentations on testing experiences and strategies, maintaining sample collections, testing heuristic capabilities, and so on.

At the AVAR conference in November, a number of presentations dealt with testing issues and, unusually, a book on malware published in September by *Syngress* [24] included chapters on malware analysis in general, and testing in particular. There was piecemeal commentary on some of the tests described above in blogs and white papers. But behind the scenes, even more interesting things were starting to happen, reflecting an interest within the research community in the rationalization and standardization of testing practices across the board, not just among the more established testing groups.

Anti-Malware Testing Standards Organization (AMTSO)

Through 2007, there were discussions between testing organizations and representatives of major players in the anti-malware industry, notably at the CARO workshop and at AVAR 2007. These discussions led directly to a two-day meeting in Bilbao in which more than 40 anti-malware specialists participated, as described at <http://www.amtso.org>.

AMTSO is intended to facilitate the improvement in the 'objectivity, quality and relevance of testing methodologies' and the establishment of standards through the development of good practice guidelines, promoting education and the provision of resources [5]. A subsequent meeting in April 2008 saw increased interest in participation from the AV community and the formalization of the organization. As the structure of the group has solidified, work has begun on a number of deliverables. It is important to note that AMTSO recognizes the possibility that it will be misperceived as an industry pressure group. It is deliberately inclusive of *all* interested parties such as

anti-malware product vendors, journalists, magazine editors, anti-malware testers and product certifiers, as well as representation from the academic community to ensure a scientifically rigorous and sound approach.

The WildCore Working Group

The WildList Organization Working Group on WildCore, Testing and Certification is a much smaller and currently less formal group. Its current membership comprises a subset of WildList Organization International (WLO or WLOI) members (including these authors). However, the group is represented within AMTSO, and the two organizations expect to work closely together within the terms of a proposed Memorandum of Understanding. Thus, the group is able to focus on a small range of issues, including some very specific to the WildList, WildCore and WLO branding, but is not isolated from the mainstream of industry and industry-independent initiatives and thinking in similar or related topic areas.

At time of writing the group consists of independent researchers and representatives of companies whose product range includes anti-malware solutions. Its general remit is to discuss and make recommendations about certification and/or comparative testing as applied to security products, especially anti-malware products and countermeasures. The group is also discussing the professionalization of the testing and certification sectors in general terms. Possible documentation deliverables address, among other issues:

- The expectations of the wider WLO community as regards general testing methodologies and best practice.
- WLO expectations and requirements regarding the use of its resources – especially WildCore, a set of replicated virus samples (<http://www.wildlist.org/faq.htm>) based on the WildList – and branding. Branding refers, in this context, less to the traditional business preoccupation with trademarks, copyright, and intellectual property rights, than it does to the misleading application of terminology associated formally or informally with WLO. For instance, the term 'in the wild' is often used in ways that don't correspond very closely to the WLO definition of 'In the Wild' (ItW) [25, 26]. This is a real problem when tests claim to cover 'in-the-wild' testing [18], but don't take into account the need for sound sample identification and validation [27].
- The provision of information on testing and other relevant methodologies. Because of the likely crossover with AMTSO deliverables, this is likely to focus on WildList (ItW) testing to avoid duplication of effort, including methodology and eligibility to access WLO resources.

Should we be certified?

A particular area of concern among some members of both groups is whether it is possible or desirable to initiate a formal process to 'certify' testers (for example, to qualify for use of/access to a 'standard' test set like WildCore), or otherwise enable them to prove their competence, knowledge, experience and ethical fitness to work in that arena. There are concerns that

such a certification framework might be initiated by a group outside the industry. There are also concerns about such a framework being controlled exclusively by the anti-malware industry *or* the anti-malware testing industry, but the potential for misinformation and misconceived certification criteria determined by organizations with inadequate understanding of (and sympathy for) anti-malware technology and ethos is particularly worrying.

Close cooperation between AMTISO and the WLO group should result in a streamlining of parallel activities, rather than reinvention of wheel families and variants. The smaller group will concentrate on WildList-specific issues in the first instance, while expanding the informational resources of which AMTISO can make use.

TESTING THE TESTERS

Good testing helps vendors to raise the quality of their products, which is not only good for the consumer but, in the long run, for the vendor's sales. However, inadequate and unfair testing misleads consumers, and may result in poor products being misrepresented as better than good products. Because of the widespread lack of comprehension of the mechanics of anti-malware technology and of the fundamentals of testing, many testers are unable to assess product effectiveness accurately [28].

How, then, do we make testers more accountable and aware of their obligation to offer scrupulously accurate information? How do we make the wider security community and the public aware of the problems caused by poor testing? Some groups are well aware of this heavy responsibility and have played a major part in the establishment of AMTISO.

Documentation offering better information on standards and on good and bad practice will help some amateur testers to test more appropriately, even though some will continue to dismiss the concerns of the industry as self-serving. However, the establishment of formal criteria for the validation of testers and testing organizations would enable some discrimination between qualified and unqualified testers.

Three distinct main areas of testing certification need to be considered:

- Firstly, 'official' certification bodies, such as *WestCoast Labs* or *ICSALabs*, though the range of certifying organizations is increasingly wide (not only in numbers, but in scope, security context, and professionalism)
- Secondly, professional testers like *AV-Test* and *AV-comparatives*.
- Thirdly, magazine reviewers.

From the tester's perspective, only members of the first category are likely to be interested in having their work officially 'standardized': indeed, they may stress their own compliance with standards such as ISO 9001 or 17025. Certification testing is characteristically a requirement for adoption by large enterprises or government bodies, who are often standards-oriented and less interested in consumer-level test results.

Members of the third category may be the group most in need of regulation, but are also the least likely to conform to the

standards of a group such as AMTISO or to accept the need to have their work certified. And, of course, there will always be individuals and groups who will see no need to conform because they don't regard themselves as professional (and indeed, may see it as intrinsically virtuous to be amateur).

It isn't practical to create a certification that all testers would be required to meet before they could publicize their tests. Those who see the relevance of certification already have some, in the form of ISO standards. Those who need it most won't necessarily accept a need for it. It may be more useful to provide resources that will help (and hopefully encourage) conformance with 'good' standards in testing performed by less able (and less well resourced) groups.

There is clearly no way to force *anyone* to use any standards whatsoever, nor would the belligerent pursuit of (for instance) journalists and editors who decide not to conform help our cause. So establishing a tester certification process would not be a trivial task. It would, perhaps, be a simpler task (at least administratively) for an organization that already offers a range of security certifications, but there is no such organization already aligned to the anti-malware industry, while such organizations within the mainstream security industry often have little technical understanding of or sympathy with the anti-malware specialist sector.

What are the issues that a certification process might (and should) address? It is probably not desirable that a body controlling the certification of testers should itself be controlled by any single sector, whether that is the academic community, the testing organizations, the anti-malware industry or their customers, all of whom have something to gain overall from the rationalization of testing processes. But what can we learn from existing bodies about the necessary components of a certification process?

- In principle, both academia and the wider security certification industry, especially organizations with detailed knowledge of ISO 17024 (assessing and certifying personnel) conformance could advise on the formal mechanism for defining what (ISC)² call a Common Body of Knowledge [29], the educational process of sharing that knowledge, and a process for assessing competence and familiarity with that material.
- Most of the major players in comparative testing and product certification testing, as well as organizations with less direct involvement such as *VirusTotal* (<http://www.virustotal.com>), and WildList Organization International (<http://www.wildlist.org>), are already closely involved with groups described above. Their experience extends beyond testing technologies to other areas such as relations with the customers who commission their services, and to compliance with existing standards such as ISO 9001 (quality management), ISO 17025 (general requirements for the competence of testing and calibration laboratories), and so on.
- The anti-malware research community is the most knowledgeable group (generally speaking) in terms of what AV technology is and does, how it works and what to expect from it, and often works cooperatively across

corporate and marketing borders. The participation of a good spread of representatives across the industry should ensure that the interests of a single vendor don't override the common good, and joint participation of other parties should ensure that the industry as a whole remains accountable to its customers. The industry also has its own experience and expertise in testing, not least in terms of Quality Assurance (QA) procedures and comparative analysis.

It is important that customer interests are also represented, for example by participation by groups such as the Anti-Virus Information Exchange Network (AVIEN), which includes major customer organizations (see <http://www.avien.net>) and already has a history of involvement with certification issues [30]. While this involvement has been focused on the projected certification of independent researchers and administrators, its framework of skills and experience is relevant to the tester of anti-malware technologies representing the interests of the consumer. A Certified Enterprise Anti-Virus Architect (CEAVA), the highest level of certification discussed by the AVIEN project team, would have skills including:

- Hands-on virus processing (recognition, isolation, replication, disassembly and analysis, submission for evaluation, and disinfection).
- Managing an AV system (installation, optimization, maintenance, troubleshooting, threat response, product evaluation, manage deployment and support).
- Developing an AV policy and strategy (creation of best practice guidelines, policy, multi-layered defensive strategy, response planning, statutory compliance management, auditing deployment and standards compliance).

There is also a proposed certification for a Certified Anti-Virus Instructor (CAVI), who would have the necessary skills and knowledge to develop and teach a live virus workshop. Clearly, many of these skills might be directly transferrable to a testing certification framework. Of course, in the real world, countermeasures against malicious code are set within the broader framework of a wider security strategy, and it is not unreasonable to expect a tester to share a similar range of experience to that of the corporate security administrator, including awareness of such areas as information governance, compliance with other standards and legislation, and so on. For tester certification purposes, the required skill set could be extended to include knowledge of (for instance):

- Formal software testing methods
- Anti-malware product test classifications
 - Comparative testing
 - Product certification
 - Single product review testing
 - Q&A testing
- Function testing types
 - Detection performance
 - o Proactive/reactive
 - o Real-time/on-demand
 - o Default/relaxed/paranoid configuration

- Other functionality
 - o Scan speed
 - o System footprint
 - o Support
 - o Documentation
 - o User interface
- Test target types:
 - WildList/zoo
 - Time to update
 - False positive (FP)
 - Dynamic analysis
 - Retrospective
 - Packer detection
 - Behaviour analysis
- Ability to explain and deploy software forensic methods, dynamic and static malware analysis, and related methodologies.

In short, the tester requires the ability to place detection and other function testing into the wider context of product evaluation while being able to discriminate between test targets so as to avoid compromising targets by introducing inappropriate or conflicting methodologies.

CONCLUSION

Some of the trouble spots we have noted over the years in comparative reviews include those shown in Table 1.

Inappropriate test sets	Test sets that include non-viruses, non-malware such as test and garbage files, and non-viable malware.
Simulated malware	This can lead to many complications: a scanner may be 'rewarded' for incorrectly diagnosing a simulation as the malware it impersonates (the purpose of a malware detector is to detect malware, and only malware).
Kit malware	Samples generated automatically by a kit are notoriously unreliable, and may not represent a real-life scenario.
Contextually inappropriate malware (or non-malware)	Examples include: <ul style="list-style-type: none"> • Testing web scanners with HTML samples that only ever appeared in the wild as SMTP transmissions [33]. • EICAR testing not distinguished from real detection testing.
Unvalidated samples	Where a sample is presumed to be malware but inadequately checked, ignoring potential false positives.
Circular validation	Malware is 'validated' by testing against one of the products under test, thus introducing a massive bias and missing potential false positives.

Apples versus oranges testing	A term frequently applied to comparative tests where products of significantly differing functionality, levels of configuration, and so on, are tested with the same test set and essential methodology without reference to their differences.
Fuzzy test targets	For example, making no clear distinction between tests of heuristic detection, generic filtering, near-exact identification, and so on.
Statistical problems	Sample size, sampling techniques, statistical bias and so on [32, 33].

Table 1: Testing trouble spots (adapted from [11]).

We have considered three main approaches to reducing the impact of such poor practices:

- Provision of better information on:
 - How (not) to test
 - How to evaluate the feasibility and accuracy of a test
 - Related issues like malware identification and naming, the ethics and mechanics of sample distribution, and so on.
- Countering misinformation and misinterpretation of test data by testers and third-party commentators.
- Increasing the accountability of the organizations responsible for product testing and certification.

The first approaches may be well covered since both AMTSO and the WildList group are expected to work on documentation-focused approaches (glossaries, FAQs, standards and guidelines). Test issues requiring consideration include retrospective/frozen/dynamic testing, time-to-update testing, false positive testing, zoo testing, speed, resource utilization, and many other variations and evaluation criteria, as well as ItW testing. FAQs and resources on other relevant topics could include:

- good test resources, tools, reading resources and information
- sample validation
- good and bad practice
- naming and sample cross-reference
- acquiring samples
- cooperation with the anti-malware industry
- measuring testing competence
- a glossary resource.

Indeed, work has already begun on some of these.

As regards the second approach, one of AMTSO’s approaches will be to ‘Provide analysis and review of current and future testing of anti-malware and related products.’ [5]. By providing a single coherent focus for reviews of reviews, tests and methodologies, the site will provide a point of reference to which the eyes of the Internet can be directed.

Finally, testers can demonstrate their awareness of the need to be accountable for the quality of their testing and the

conclusions drawn from their test data by going through a certification process to confirm their credibility and competence. Intermediate, shorter-term measures are also possible. For example, AMTSO or an approved alternative group might offer the opportunity for a testing organization to validate its testing processes through external audit, or through a less rigorous checklist of criteria to apply, or even an expressed intent to comply with AMTSO recommendations. While some testers will be disadvantaged by an inability or reluctance to pay for expensive subscriptions, auditing and so on, even awareness of the issues addressed by AMTSO would make a favourable impression.

Sound testers, the industry and academia cannot enforce universal standards without the support of a considerably wider community. However, if the organization makes better information available, that’s a quick and easy win. A partial solution is still better than no solution even though poor testing will continue and misinformation will flourish.

One thing AMTSO could certainly do is propose a generic ethical code or framework for testing. For example, there is a need to raise awareness of tester accountability. A tester is accountable to his or her audience, not to the providers of tested products or services. This is not a get-out-of-jail-free card, but the opposite. The tester should represent accurately the value (or otherwise) of the product/service to the audience. If a test misrepresents – deliberately or otherwise – the value or functionality of the product, it doesn’t meet the standards of ethical behaviour expected (implicitly) by the audience. Whether or not test results are marketed commercially, a service is offered to an audience, which is entitled to expect adherence to reasonable standards of truth, accuracy, and ethical and moral behaviour (unless a tester confesses to being dishonest and incompetent!).

AMTSO is an expression of the need for clear standards, but also of a way to communicate those standards as achievable goals for those aspiring testers. While there have been objections on the fringes of the anti-malware community [34] to the ‘unscientific’ basis of the AMTSO initiative, we believe that the expertise of an experienced industry, the representation within AMTSO of the academic community combined with ongoing discussion and an imposing, already existing corpus of published work can make a significant impact on a problem that has plagued us for long enough.

REFERENCES

[1] Harley, D. I’m OK, you’re not OK. Virus Bulletin, November 2006. <http://www.virusbtn.com/virusbulletin/archive/2006/11/vb200611-OK>.

[2] Lee, A. Testing heuristic detection in a real-world scenario. 2004. [http://www.eset.com/download/whitepapers/Eset_March06draft%20\(2\).pdf](http://www.eset.com/download/whitepapers/Eset_March06draft%20(2).pdf).

[3] Harley, D.; Lee, A. Testing, testing: anti-malware evaluation for the enterprise. Proceedings of the 10th Association of anti Virus Asia Researchers International Conference, 2007.

[4] <http://www.f-prot.com/workshop2007/presentations.html>.

- [5] Anti-Malware Testing Standards Organization. AMTSO Formation Press Release. http://www.amtso.org/index.php?view=article&catid=2%3Apressreleases&id=5%3Aformationpressrelease&option=com_content&Itemid=2.
- [6] Wells, J. Open letter. <http://www.cyber.com/details.php?id=14§ion=detailpapers>.
- [7] Consumer Reports. <http://www.consumersearch.com/www/software/antivirus-software/review10271.html>.
- [8] Morris, D. AntiVirus FightClub Results! <http://blog.untangle.com/?p=96>.
- [9] Paller, A. Consumer Reports creates 5,500 viruses for tests. <http://www.sans.org/newsletters/newsbites/newsbites.php?vol=8&issue=65&rsID320>.
- [10] Harley, D. Untangling the wheat from the chaff in comparative anti-virus reviews. http://www.smallblue-greenworld.co.uk/AV_comparative_guide.pdf.
- [11] Harley, D. AV testing SANS virus creation. Virus Bulletin, October 2006. <http://www.virusbtn.com/virusbulletin/archive/2006/10/vb200610-sans>.
- [12] Reader comments on <http://www.avertlabs.com/research/blog/index.php/2007/08/12/what-a-tangled-web/>.
- [13] Reader comments on http://blog.washingtonpost.com/securityfix/2006/08/antivirus_testing_and_consumer_1.html.
- [14] Abrams, R. You have to try hard to be less competent. <http://www.eset.com/threat-center/blog/?p=78>.
- [15] Dunn, J. Consumer group slammed for creating ‘test’ viruses – ‘Why would anyone ... want to add to the glut?’. http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9002499&source=rss_topic17.
- [16] Hayter, A. Nature of anti-malware testing and certification programs. Proceedings of the 10th Association of anti-Virus Asia Researchers International Conference, 2007.
- [17] Morgenstern, M.; Marx, A. Testing of ‘Dynamic Detection’. Proceedings of the 10th Association of anti Virus Asia Researchers International Conference, 2007.
- [18] Morris, D. AntiVirus FightClub conclusions. <http://blog.untangle.com/?p=99>.
- [19] Heise Security. Antivirus protection worse than a year ago. <http://www.heise-online.co.uk/security/Antivirus-protection-worse-than-a-year-ago--/news/100900>.
- [20] Landesman, M. Testing hocus pocus. <http://antivirus.about.com/b/2006/08/17/testing-hocus-pocus.htm>.
- [21] Seltzer, L. Ethics and virus testing. <http://www.eweek.com/c/a/Security/Ethics-and-Virus-Testing/>.
- [22] Schmidt, J. Thou shalt not create new viruses. <http://www.heise-online.co.uk/security/Is-it-permissible-to-create-new-viruses--/features/77440>.
- [23] Eckleberry, A. More testing silliness. <http://sunbeltblog.blogspot.com/2006/08/more-testing-silliness.html>.
- [24] Harley, D. (ed.). AVIEN Malware Defense guide for the Enterprise. Syngress, 2007.
- [25] WildList Organization International. What exactly is ‘In the Wild’? <http://www.wildlist.org/faq.htm>.
- [26] Ducklin, P. Counting viruses. Proceedings of the 9th International Virus Bulletin Conference. 1999.
- [27] Harley, D.; Lee, A. Antimalware evaluation and testing. In Harley, D. (ed.) AVIEN Malware Defense Guide for the Enterprise (Syngress 2007).
- [28] Harley, D. Security zone: the trouble with testing anti-malware. <http://www.computerweekly.com/Articles/2008/01/07/228766/security-zone-the-trouble-with-testing-anti-malware.htm>.
- [29] The International Information Systems Security Certification Consortium, Inc. About the (ISC)² CBK. <https://www.isc2.org/cgi-bin/content.cgi?category=7>.
- [30] Bechtel, K.; Harley, D. Customer power and AV wannabes. In Harley D. (ed.) AVIEN Malware Defense Guide for the Enterprise. Syngress 2007.
- [31] Muttik, I. A tangled web. In Harley D. (ed.) AVIEN Malware Defense Guide for the Enterprise. Syngress, 2007.
- [32] Dang, H. What a ‘Tangled’ Web... <http://www.avertlabs.com/research/blog/index.php/2007/08/12/what-a-tangled-web/>.
- [33] Muttik, I. Comparing the comparatives. Proceedings of the 11th Virus Bulletin International Conference, 2001. http://www.mcafee.com/us/local_content/white_papers/threat_center/wp_imuttik_vb_conf_2001.pdf.
- [34] Hayter, A. Report from the 17th EICAR Annual Conference. <http://www.aavar.org/'08%20eicar%20report%202.html>.

APPENDIX 1: TOPICS FOR A CERTIFICATION FRAMEWORK

Here are some of the areas that need to be understood by a tester, and therefore considered as possible topics within a certification framework, in addition to the extended skill set described above:

- **Test transparency:** Reproducibility in detection testing is difficult to achieve. A snapshot test at one moment in time may not be reproducible later because the currency of the sample set, the detection status of the tested product (as regards heuristic rule updates, signature updates, engine updates and so forth), and the fine detail of the methodology and test rig are all liable to change very quickly, even if testers are prepared to share resource

information. Nonetheless, the tester must give enough information about the test (sample sets, product versions and configuration, methodology, etc.) to allow the audience to assess the validity of the test. (This does pose questions as to how much information is enough and how much reliance can be placed on the ability of the audience to make an assessment, irrespective of the quality of the available information.) Ideally, this topic should also include consideration of the extent to which the audience and the suppliers of the tested products can actively discuss the test with the tester before, during and after testing.

- **Statistical issues:** sample rightsizing, sampling techniques, metrication and instrumentation, realistic analysis, bias exclusion and so on.
- **Ethical issues**, including:
 - Responsible disclosure (results, problems encountered, vulnerabilities and exploits unearthed)
 - Right to preview and right to reply
 - Sharing of samples
 - Declaration of stakeholding and potential conflicts of interest
 - Generation of sample sets (ethical sourcing, issues around virus and/or malware creation)
 - Duty of care (safety issues)
 - Responsibility (perhaps the equivalent to the medical practitioner's 'Do no harm' is 'Do not mislead'?)
- **Sound methodologies:**
 - Apples and apples, not apples and oranges
 - Platform consistency
 - Consistency of test objectives over the course of the test
 - Selection of appropriate test scenarios and sample sets: application, measurement, calculation, analysis and interpretation
- **Standards:** Formulation, establishment, teaching and understanding of standards and guidelines (plus relevant legislation, objective standards such as BSI and ISO standards, and so on), adherence to which is in itself a measure of competence.
- **Understanding of keywords and key concepts:**
 - Objectivity
 - Currency
 - Validation
 - Verification
 - Reproducibility
 - Consistency
 - Targeting
 - Discrimination
 - Clarity of Objectives
- **Metrics:** measuring the efficacy of protection (effectiveness, security). This actually addresses a very

subjective yardstick: protection is a continuum between convenience prioritized over security, and security at the expense of convenience.

APPENDIX 2: ISO QUALITY STANDARDS

ISO 9000 is a family of international standards for quality management systems, described in ISO 9000:2005. It covers such requirements as monitoring processes for effectiveness, and adequate record-keeping, quality assessment and monitoring. Certification is not intended to prove the quality of products and services, but that appropriate and consistent processes are in place.

See: http://www.iso.org/iso/catalogue_detail?csnumber=42180.

ISO/IEC 17025 is an international standard covering the requirements that must be met to prove competence to carry out tests and/or calibrations, using standard, non-standard and lab-developed methods. It is described in ISO/IEC 17025:2005, and intended to be used to develop quality, administrative and technical operations.

See: http://www.iso.org/iso/catalogue_detail?csnumber=39883.