# Antivirus Testing and AMTSO
## Has anything changed?

**David Harley BA CITP FBCS CISSP**
**ESET Senior Research Fellow**

Cyber Threat Analysis Center
ESET LLC, San Diego, CA 92101, US
Email: david.harley@eset.com

**CFET 2010**
**4th International Conference**
**on Cybercrime Forensics Education and Training**

# Abstract

*Since it was formally founded in May 2008, the Anti-Malware Testing Standards Organization has been through a number of changes and generated some serious documentation and significant press coverage. AMTSO was actually founded as the result of many years of concern, not to say rage on occasion, on the part of anti-malware vendors and mainstream product testers, at the low level of competence and accuracy demonstrated by so many of the individuals and organizations offering comparative testing and/or product certification.*

*The organization announced its intention of improving levels of objectivity, quality and relevance of anti-malware testing methodologies. Clearly that wasn't going to happen overnight, but how far along the road to better testing practice have we travelled?*

*This paper looks at testing as it was, as it is, and as AMTSO would like it to be. Is testing really so difficult? Is it appropriate for the vendors who make the products under test to be so involved in the process of defining good practice? In the process, core issues will be considered such as:*

- *Comparative testing versus certification*
- *Detection testing versus performance testing, and why it's rarely a good idea to mix the two*
- *Detection testing in a time of glut: when a virus lab may process tens or hundreds of thousands of unique binaries on a daily basis, prioritization is not a trivial issue. How big is the margin for error?*
- *Comparing apples to oranges: can you penalize an orange for not tasting like an apple?*
- *Default configuration and level playing fields*
- *Correct classification and selection of samples.*
- *Validation: is that sample really malicious, and how does a tester check?*
- *Static analysis and static testing: is there still a place for signatures and WildList testing?*
- *Is a good static test better than a bad dynamic test?*
- *The AMTSO fundamental principles of testing: do they help or hinder? Is standardization of testing even a good idea?*

# 1.0 Introduction

Once upon a time, anti-malware testing was an absolute free-for-all. Three blokes could get together in a bar (Harley & Lee 2008) and say, "I know: let's start testing anti-virus products. We could even charge for it." Forums like alt.comp.virus were regularly visited by individuals asking to be given virus samples so that they could do some product testing, among other reasons...

Since testers were not required to validate their own qualifications, the tools they used were many and various, and subject to no real controls. Journalists would review products based on their own testing with:

- Sample sets they'd picked up somewhere on the internet, possibly "validated" by their own favourite scanner (often a particular open source scanner that didn't cost them anything), or by submitting samples to VirusTotal, and often with no validation at all, freeing them of the need to worry about false positives, garbage files and so on (Harley 2007)
- Samples supplied by one of the products they were testing (strangely enough, those products usually did exceptionally well)
- Simulated viruses (Wells et al 2000; Gordon 1995)
- Kit-generated viruses (Dunn 2008)]
- "Virus-like" files (whatever that means...) that they'd created themselves
- Newly-created malware, or some facsimile thereof
- Modified versions of real malware
- Modified versions of the EICAR test file (Dechaux et al. 2010)

While the intentions of many testers may have been good, approaches like this are at best bound to generate some problems, and in some cases are totally inappropriate and misleading, and open to all kinds of abuse (Tanner 1993). While there were occasional concerted efforts to respond to a particularly inappropriate test [Wells 2000], the overall impression in the public mind was of a peevish antivirus industry that didn't like the way testing was being carried out in general, but was reluctant to provide positive feedback on how testers could improve (Harley & Lee 2008).

In fact, the WildList Organization (http://www.wildlist.org) developed in part as a way to enable testers to improve testing with the use of a validated set of known In-the-Wild malware (Harley & Lee 2010). In the meantime, a handful of highly professional testers evolved ways to work in partnership with AV vendors while preserving their independence, but these activities did not generally impact upon the consciousness of the general public.

# 2.0 AMTSO and After

The Anti-Malware Testing Standards Organization was formally founded in May 2008 (AMTSO 2008), though the meeting in Bilbao which resulted in that formalization was the culmination of years of discontent (Harley & Lee 2007), as described above, and some serious discussion between some vendors and mainstream testers. Perhaps the first indication that this discussion was likely to inspire serious changes in the testing landscape can be found in a group of presentations (CARO 2007) made at the International Antivirus Testing Workshop 2007, held at Reykjavik, Iceland under the auspices

of CARO (Computer Anti-virus Researchers Organization), and a cluster of presentations (Harley & Lee, 2007) and informal discussions at AVAR later the same year continued the theme.

## 2.1 The AMTSO Charter

Following the 2008 meeting, the organization announced its intention of improving levels of objectivity, quality and relevance of anti-malware testing methodologies (Table 1) in its charter.

| | |
|---|---|
| Item 1 | Providing a forum for discussions related to the testing of anti-malware and related products. |
| Item 2 | Developing and publicizing objective standards and best practices for testing of anti-malware and related products. |
| Item 3 | Promoting education and awareness of issues related to the testing of anti-malware and related products. |
| Item 4 | Providing tools and resources to aid standards-based testing methodologies. |
| Item 5 | Providing analysis and review of current and future testing of anti-malware and related products. |

**Table 1: AMTSO Charter**

Clearly that wasn't going to happen overnight (Harley 2010), but how far along the road to better testing practice have we travelled? Clearly, bad (or even contentious) testing hasn't gone away, and three blokes in a bar can still announce that they are now a testing or certification organization: so what has been achieved?

## 2.2 Deliverables

Certainly, some deliverables are easily enumerated, such as the documentation approved (Table 2) to date by the AMTSO membership at the organization's workshops.

| Item | Document Name | Date Approved |
|---|---|---|
| 1 | AMTSO Fundamental Principles of Testing | 31/10/2008. |
| 2 | AMTSO Best Practices for Dynamic Testing | 31/10/2008 |
| 3 | AMTSO Best Practices for validation of samples | 7/5/2009 |
| 4 | AMTSO Best Practices for Testing In-the-Cloud Security Products | 7/5/2009 |
| 5 | AMTSO Analysis of Reviews Process | 7/5/2009 |
| 6 | AMTSO Guidelines for testing Network Based Security Products | 13/10/2009 |
| 7 | AMTSO Issues involved in the "creation" of samples for testing | 13/10/2009 |
| 8 | AMTSO Whole Product Testing Guidelines | 25/5/2010 |
| 9 | AMTSO Performance Testing Guidelines | 25/5/2010 |

**Table 2: Documentation Deliverables**

The deliverables in Table 2 cover a wide range of technical issues as well as (especially in the case of item 7) ethical issues. Item 1, however, is, while not particularly technical in itself, particularly significant in that it provides a high-level view of the principles that underpin the other documents, and arguably AMTSO's raison(s) d'être.

## 2.3 Fundamental Principles

As the nine principles enumerated in Item 1 have generated much of the controversy considered later, they are listed in Table 3 below.

| | |
|---|---|
| 1 | Testing must not endanger the public |
| 2 | Testing must be unbiased. |
| 3 | Testing should be reasonably open and transparent |
| 4 | The effectiveness and performance of anti-malware products must be measured in a balanced way |
| 5 | Testers must take reasonable care to validate whether test samples or test cases have been accurately classified as malicious, innocent or invalid |
| 6 | Testing methodology must be consistent with the testing purpose |
| 7 | The conclusions of a test must be based on the test results |
| 8 | Test results should be statistically valid |
| 9 | Vendors, testers and publishers must have an active contact point for testing-related correspondence |

**Table 3: The Nine Principles**

## 2.4 Resources

The Resources page at http://www.amtso.org/en/related-resources.html started as an aggregation of external links to specialist testing-related resources. However, as the organization has started to gather momentum, it has been extended to include resources generated in some sense under the auspices of AMTSO, from "here we are" briefing material explaining the provenance and aims of the organization (Harley D. 2010a) to technical material from more recent conferences.

## 2.5 Making Testers Accountable

No-one in the anti-malware industry believes that they aren't accountable to their customers. Perhaps, indeed, the industry is less effective in terms of absolute security because it tries to meet all the expectations of its customers, realistic or otherwise (Harley 2006).

Is the anti-malware industry accountable to the testing industry? Yes, in so far as the testing industry has some claim (implicit or otherwise) to represent the interests of customers. But should the testing industry be directly accountable to the anti-malware industry? Clearly, that would be too open to abuse,

Nonetheless, testers, journalists and publishers should clearly be accountable to their audiences for the accuracy and relevance of their conclusions. Sharing information and samples not only helps vendors to improve *their* products, but allows testers to improve *their* testing by introducing a check on its quality, and demonstrates their awareness of the need to test responsibly (Harley 2010b).

However, some testers are unable to share information because of internal policy, or because they regard information sharing as a threat to their impartiality and independence. Others are hampered by legal and disclosure considerations. However compelling, these issues compromise the perceived validity of a test if there are no checks on methodology, and therefore on whether the published conclusions follow logically from the methodology and the data.

Lack of direct accountability to the providers of tested products or services 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, the tester is offering a service to an audience, which is entitled to expect adherence to reasonable standards of truth, accuracy, and ethical and moral behaviour.

## 3.0 Comparative Testing

Comparative detection testing, where the tester attempts to rank products according to their effectiveness in terms of detection malware is based on a number of assumptions that don't always hold up (Harley & Lee 2010). Most notably, the contention that testers are better at gathering, validating and classifying samples than vendors. The same doubts apply in terms of performance testing, by which I mean "performance" in areas other than detection testing such as resource usage, memory footprint and scanning speed. While you can certainly talk about "detection performance", in technical discussion it's been found more useful to separate consideration of detection based on raw detection scores from more general performance issues such as ergonomic feasibility, impact on system performance, and usability (Vrabec & Harley 2010). The intention, however, is not to understate the importance (or difficulty) [performance testing paper] of general performance testing (AMTSO 2010).

Since it isn't really feasible to use the whole 40 million or so known samples of malware in detection testing (as of May 2010, according to several presentations at the CARO workshop that took place in Helsinki during that month (Sophos 2010) a set of test samples is presumably intended to represent an accurate reflection of the totality of malicious programs currently in the world. However, that accuracy and relevance is pretty difficult to quantify (Kuo 2009), though post-test validation of test samples by vendors indicates an uncomfortably wide margin for error (Košinár et al. 2010).

At best, we have to accept testing on the basis of professionalism, best endeavours and reasonable competence rather than exact quantification. The AMTSO review analysis process cannot be regarded as a substitute for the independent certification of testers and/or testing organizations by an appropriate agency, but it's hard to see how an organization in which vendors are so strongly represented could, by itself, provide such a credibly independent certification. The review analysis process does, however, at least provide a means to assess whether a given test or test report meets those standards by comparing it to the AMTSO "Fundamental Principles".

## 3.1 Certification Testing

Certification testing as applied to products isn't based on finding the "best" product, magnifying small variations in performance (detection *or* general performance), or isolating layers of protection artificially, in order to establish clear winners and losers. Rather, it's based on establishing a baseline value for consistently acceptable performance, and that tends to result in a premium being placed on the ability to validate, rather than on pushing the detection envelope, which is why WildList testing still maintains a (diminishing) presence in this area of testing (Harley & Lee 2010), though it's unusual nowadays for a certification test not to take into account other forms of testing more in line with the dynamic, real-world tests favoured by AMTSO.

A score of 100% in a WildCore-based test such as VB100 does *not* imply 100% detection of all realworld threats or anything near it. It has some residual value in certification testing, less in comparative testing. That value lies in the fact that it can be verified. The kind of dynamic (in a broad sense) testing that AMTSO members and others have been advocating (Muttik & Vignoles 2008) is much more difficult to execute and to verify, and that's the sort of problem that AMTSO members are trying to solve. In the meantime, WildLost testing still tells you something useful. It doesn't tell you which is the "best" product at overall detection, but nor does a comparative test that offers no way to validate its methodology or test set.

In fact, a certification test may be about testing a number of other aspects of detection and performance (Harley 2009). Testing multiple layers of detection and other aspects of functionality provide a more complete picture of the capabilities of a product than a single limited detection test, especially a static test. However, it's rarely a good idea (Harley 2007; Vrabec 2010) to mix detection testing and performance testing in the same test. For example, where a product is optimized for speed by default, it may miss samples it is quite capable of detecting in a more paranoid mode. That's potentially a problem in comparative and in certification testing, and also illustrates a persistent problem in testing where the execution of the test relies on the use of default settings.

## 3.2 Testing By Default

Testing based purely on default configuration is often defended on the grounds that it matches the most common user experience, since received wisdom is that in general, end users do not change configurations. In fact, there's nothing wrong in principle with a reviewer examining a product's performance in a default configuration, as long as the tester realizes that he's not actually testing detection, but design philosophy, and makes that clear to his audience. But security on any given system is not an absolute: it's a locus on a continuum between total security and total convenience. That locus represents a choice on the part of the user to use or modify the default configuration. Either way, if the user is a tester, both he and his audience should be aware that he's testing only a subset of the product's functionality. The tester has, therefore, a responsibility to ensure that the test offers a fair comparison between comparable products, in comparable configurations (Harley 2009a).

Similarly, differences in the ways that scanners regard edge cases such as Possibly Unwanted applications may have knock-on effects on apparent detection: while some scanners treat "greyware" as malware by default, others are more cautious. While this is clearly an indication of the need to consider carefully whether a default configuration is appropriate to the test, it's also an illustration of the need to validate and classify samples correctly. (Harley, 2009a)

## 3.3 Static versus Dyamic

Unfortunately, these terms are used in such a wide variety of contexts, even within the context of anti-malware technology,

Let's take two seconds to look at how most AV scanners work:

- Passive scanning is fairly quick and easy to test on-demand: on-access (real-time) scanning requires more resources and effort to test with the same size of sample set, but it is at least feasible to automate the process, time constraints allowing. It's roughly equivalent to static (code) analysis as the term is used in digital forensics.
- Active/Dynamic Scanning is roughly equivalent to dynamic analysis in forensics, using allied techniques such as emulation, virtual machines, and sandboxing to implement some form of behaviour analysis. However, it introduces a processing delay and is platform-specific: that is, it may, according to product, be more restricted in capability and may not even be implemented on all platforms.

Static testing is often based on on-demand scanning, where the security software is allowed to scan a file/object without allowing it to execute, though it may execute in a virtual environment (emulation), allowing the product to make use of some form of behaviour analysis. However, this approach is highly product-specific: the fact that a threat isn't detected by a given product under these circumstances doesn't mean that it isn't capable of detecting it in a different execution context (Harley, 2009). But it does have advantages for the tester:

- Very convenient testing practice.
- Can be done simply by running on-demand, even command-line scanner.
- Almost platform independent, if detection database is standard across platforms.

Sometimes, however, an on-demand scanner doesn't pick up what an on-access scan does. Command-line scan tests are particularly likely to penalize products that use active heuristics or another form of behaviour analysis, especially if they're not running on the platform to which the malware is native – say a Linux scanner checking an obfuscated Windows Trojan.

If there is no execution there is no behaviour to observe/analyse. So the result doesn't reflect real-world, real-time detection for a product that doesn't use behaviour analysis in an on-demand scan. It's for these reasons that AMTSO guidelines documentation has focused on dynamic testing (in a broad sense) by addressing such issues as dynamic testing, In-the-Cloud testing, and network product testing (AMTSO 2010). Still, it can be argued that a good static test is better than a bad dynamic test, within limits.

The AMTSO principles aren't inconsistent with good static testing, even though AMTSO does, in general, advocate dynamic testing as the way to go for testing that reflects the real world reasonably accurately, and AMTSO guidelines documents mostly relate to dynamic testing, live network testing and so on. How can that be? Because AMTSO testing isn't about standardized testing methodologies, but about high-level principles relating to accuracy, relevance, transparency and so on. A good static test should leave its audience reasonably well-equipped to appreciate its

limitations. Compare that to a dynamic test where the methodology is black box and the samples aren't shared, so all the audience has is the testing organization's assurance that its methodology is totally correct...

## 4.0 What Do You Think Of It So Far?

Well, AMTSO hasn't eliminated bad testing. Hopefully, it's published enough information to make it easier to assess good versus bad testing, but there are no instant checklists, and the review analysis process is slow to implement and remains controversial: if anything, it has reinforced the popular prejudice against the anti-virus industry. Suspicions remain that AMTSO, despite the participation of a majority of mainstream, specialist testing organizations, is somehow weighting tests in favour of technologies that suit its own sinister purposes. Perhaps the current polarization of views (AMTSO 2010) on the usefulness or otherwise of the organization is a necessary adjunct of its dramatic increase in public visibility,

### 4.1 Standards versus Guidelines

So, I asked earlier: do the AMTSO "fundamental principles of testing" help or hinder? Well, it's hard to argue with transparency, relevance and lack of bias. But is standardization of testing even a good idea? (EICAR 2010) Certainly, some testers have expressed a fear that it would compromise their ability to provide good testing. But perhaps the use of the word "standards" in the organization's name does it no favours here.

AMTSO doesn't set standards in a formal sense like BSI or ISO (2010b) and does not say who is or isn't allowed to test. Perhaps someone should, but it would not be appropriate for a body controlling the certification of testers to be controlled itself by any single sector, whether that is the academic community, the testing organizations, the anti-malware industry or their customers (Harley & Lee 2008). Furthermore, AMTSO does not prescribe testing methodologies: rather, it provides guidance at varying levels of technical sophistication, put together and approved by people with considerable expertise in complementary aspects of testing and the technology under test.

As Andrew Lee has commented, "AMTSO is not about dictating truth, but rather pointing out ways in which truth can be reliably found (and importantly, where it cannot)" (Lee 2010).

Journalist Kevin Townsend asked a number of questions about AMTSO that give some idea of the doubts that have been expressed about its provenance.

> Is this the anti-malware industry looking after itself? (It seems to be almost entirely composed of anti-malware companies and anti-malware testing companies; with little if any input from users.)

Well, few are naive enough to assume that the anti-malware vendors aren't interested in their own bottom line. As I responded to that question:

> All testing hurts products that get bad reviews. But it's not *only* about marketing and sales. Poor testing is at best irrelevant but when testing that hurts good products while promoting bad products is not *only* bad for the misevaluated product. It's much worse for the customer who puts his or her trust in a product that gives less protection than a test suggests.

## 5.0 Conclusions

Testing any software is much harder than most people seem to think it is, and testing anti-malware products is particularly hard, since few people are well acquainted with the esoteric of malicious and security technologies. Detection is obviously important in evaluating an anti-malware scanner, but it's particularly hard to test accurately. The idea that anyone with some scanners and some samples can do an adequate test doesn't hold up to scrutiny. Even the best testers can produce significantly inconsistent results, and that's inevitable given the nature of the threat landscape.

Dynamic testing is, in principle, a better reflection of "real life" testing than static testing . However, it's expensive and resource intensive, and appropriate methodologies are evolving.

Anyone can claim to be doing competent, impartial testing. It may be naive to take their word for it, especially if they're evasive about sharing methodological information and samples.

A lot of amateur testing and some for-fee testing is apples to oranges, trying to compare products that aren't intended to work in the same way. Comparing apples to apples is another issue, but you still have to configure products in the same way in order to compare them with reasonable accuracy. Out-of-the-box testing is not a level playing field unless all you're testing is out-of-the-box configuration. Even then, a review based on such testing is likely to reflect the prejudices of the tester better than the overall capabilities of the products, or even its detection capability.

AMTSO is important because it pools knowledge from the testing industry and the security industry, and each learns from and restrains the other, implementing a functional system of checks and balances. AMTSO isn't just an AV pressure group, or the AV's way of keeping the testers in line. It's better seen as an educational resource and discussion forum. Its ultimate role may be to coordinate rather than to enforce, improving methodology by enhancing objectivity, quality and relevance.

## References


AMTSO, AMTSO in the Media, http://amtso.wordpress.com/amtso-in-the-media/, 2010

AMTSO, Performance Testing Guidelines, http://www.amtso.org/amtso---download---amtso-performance-guidelines.html, 2008

AMTSO, Security Software Industry Takes First Steps Towards Forming Anti-Malware Testing Standards Organization, http://amtso.org/amtso-formation-press-release.html, 2008

AMTSO, Documents and Principles, http://amtso.org/documents.html, 2008-2010

CARO, http://www.f-prot.com/workshop2007/presentations.html, 2007

Dechaux, J., Fizaine, J., Griveau, R., & Jaafar, K., New trends in Malware Sample-Independent AV Evaluation Techniques with Respect to Document Malware, 19th EICAR Annual Conference Proceedings, ESEIA, p 93-114, 2010

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, 2008

EICAR, ICT Security: Quo Vadis? http://www.eicar.org/conference/, 2010

Gordon, S., Are Virus Simulators Still A Good Idea? http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VJG-3WP2C4W-4&_user=10&_coverDate=09%2F30%2F1996&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_searchStrId=1397764983&_rerunOrigin=google&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=ce3def7e5f12cb12f560a468b02c761d, 1995

Harley, D. & Lee, A., Call of the WildList: Last Orders for WildCore-Based Testing? Virus Bulletin Conference Proceedings, 2010 (in press)

Harley, D. & Lee, A., Testing, Testing: Anti-Malware Evaluation for the Enterprise, http://www.eset.com/resources/white-papers/Testing_Testing.pdf 2007

Harley, D. & Lee, A., Who Will Test The Testers?, http://www.eset.com/resources/white-papers/Harley-Lee-VB2008.pdf, 2008

Harley, D. AMTSO Inside and Outside, http://blogs.securiteam.com/index.php/archives/1399, 2010

Harley, D., AMTSOlutely Fabulous, Virus Bulletin p. 11-12, January 2010a

Harley, D., AMTSO not ISO, http://amtso.wordpress.com/2010/07/06/amtso-not-iso-standards-and-accountability/, 2010b

Harley D., Execution Context In Anti-Malware Testing, http://smallbluegreenblog.wordpress.com/2009/05/15/execution-context-in-anti-malware-testing/, 2009

Harley, D., I'm OK You're Not OK, www.virusbtn.com/virusbulletin/archive/2006/11/vb200611, 2006

Harley, D., Making Sense of Anti-Malware Comparative Testing, http://dx.doi.org/10.1016/j.istr.2009.03.002, 2009a

Harley, D. Untangling the Wheat from the Chaff in Comparative Anti-Virus Reviews, http://www.eset.com/resources/white-papers/AV_comparative_guide.pdf, 2007

Košinár, P., Malcho, J., Marko, R., Harley, D., AV Testing Exposed, Virus Bulletin Conference Proceedings, 2010 (in press)

Kuo, J., Let Telemetry Be Your Guide, http://blogs.technet.com/b/mmpc/archive/2009/07/16/let-telemetry-be-your-guide-a-proposal-for-security-tests.aspx, 2009

Lee, A., The edge of reason(ableness): AV Testing and the new creation scientists, http://avien.net/blog/?p=539, 2010.

Muttik, I. & Vignoles, J., Rebuilding Anti-Malware Testing for the Future, http://www.mcafee.com/us/local_content/misc/dec09.pdf, 2008.

Sophos, CARO Workshop 2010 - Day Two, http://www.sophos.com/blogs/sophoslabs/?p=9750, 2010

Tanner, S. A Reader's Guide to Reviews, "Virus News International", November, pp. 40-41, 48. 1993

Townsend, K. AMTSO: a serious attempt to clean up anti-malware testing; or just a great big con? http://kevtownsend.wordpress.com/2010/06/15/amtso-a-serious-attempt-to-clean-up-anti-malware-testing-or-just-a-great-big-con/, 2010.

Vrabec, J. Generalist Anti-Malware Product Testing, http://www.eset.com/blog/2010/01/25/generalist-anti-malware-product-testing 2010.

Vrabec, J. & Harley, D., Real Performance?, http://amtso.org/uploads/eicar2010-harley-realperformance.pdf 2010.

Wells, J. et al, Open Letter, http://www.cybersoft.com/whitepapers/paper_details.php?id=14, 2000