After AMTSO
A Funny Thing Happened On The Way To The Forum

David Harley
ESET N. America

About Author(s)

David Harley CITP FBCS CISSP has been researching and writing about security since 1989, and has worked with ESET North America – where he holds the position of Senior Research Fellow – since 2006. He previously managed the UK's National Health Service Threat Assessment Centre and is CEO of Small Blue-Green World. He is a former director of AMTSO. His books include Viruses Revealed and The AVIEN Malware Defense Guide for the Enterprise. He is a prolific writer of blogs, articles and conference papers. He is a Fellow of the BCS Institute (formerly the British Computing Society, and has held qualifications in security management, service management (ITIL), medical informatics and security audit.
Contact Details: c/o ESET North America, 610 West Ash Street, Suite 1700, San Diego, CA 92101, USA, phone +1-619-876-5458, e-mail david.harley@eset.com

Keywords

AMTSO, anti-malware, antivirus, product testing, detection testing, vested interests, vendors, testers, testing standards, testing guidelines, credibility gap, Anti-Malware Testing Standards Organization

After AMTSO
A Funny Thing Happened On The Way To The Forum

Abstract

Imagine a world where security product testing is really, really useful.

- *Testers have to prove that they know what they're doing before anyone is allowed to draw conclusions on their results in a published review.*
- *Vendors are not able to game the system by submitting samples that their competitors are unlikely to have seen, or to buy their way to the top of the rankings by heavy investment in advertising with the reviewing publication, or by engaging the testing organization for consultancy.*
- *Publishers acknowledge that their responsibility to their readers means that the claims they make for tests they sponsor should be realistic, relative to the resources they are able to put into them.*
- *Vendors don't try to pressure testers into improving their results by threatening to report them to AMTSO.*
- *Testers have found a balance between avoiding being unduly influenced by vendors on one hand and ignoring informed and informative input from vendors on the other.*
- *Vendors don't waste time they could be spending on enhancing their functionality, on tweaking their engines to perform optimally in unrealistic tests.*
- *Reviewers don't magnify insignificant differences in test performance between products by camouflaging a tiny sample set by using percentages, suggesting that a product that detects ten out of ten samples is 10% better than a product that only detects nine.*
- *Vendors don't use tests they know to be unsound to market their products because they happened to score highly.*
- *Testers don't encourage their audiences to think that they know more about validating and classifying malware than vendors.*
- *Vendors and testers actually respect each others work.*

When I snap your fingers, you will wake out of your trance, and we will consider how we could actually bring about this happy state of affairs. For a while, it looked as if AMTSO, the Anti-Malware Testing Standards Organization, might be the key (or at any rate one of the keys), and we will summarize the not inconsiderable difference that AMTSO has made to the testing landscape. However, it's clear that the organization has no magic wand and a serious credibility problem, so it isn't going to save the world (or the internet) all on its own. So where do we (the testing and anti-malware communities) go from here? Can we identify the other players in this arena and engage with them usefully and appropriately?

Introduction

AMTSO started out from what looked like (AMTSO, 2008a) a very positive place, combining the expertise of some of the best testers and that of the people who should know most about the inner workings of malware and anti-malware (the AV industry!), all of whom felt that traditional static testing methods no longer give a fair assessment of product capabilities, assuming they ever did.

Unfortunately, mistrust of the AV industry has also proved a constant barrier to AMTSO's attempts to raise the quality of testing. As AMTSO gained media attention and almost simultaneous criticism

(constructive and otherwise), the fact that it included both vendors and testers invited suspicion (Harley, 2010a). In fact, anti-malware companies do a fair amount of testing themselves, or commission it from testing organizations - not just QA and such, but comparative testing, though that sort of analysis isn't necessarily made public and tends to be (not unsurprisingly and often quite rightly) distrusted when it is.

Historically, the relationship between tester and vendor has always been complex and sometimes tense (Harley & Bridwell, 2011).

- Testers and vendors have access to some of the same sample and URL resources, as well as their own honeytraps and honeynets, samples submitted direct from the public and so on. However, the reliability of such sources is highly dependent on the quality of verification both at source and subsequently during the test process: unless there's good communication between tester and vendor, that's a significant potential stress point. Some testers verify samples with vendors before publication or at least allow some right of reply following publication. That's in accordance with AMTSO principles (AMTSO, 2008b) 3, 5, and 9, which is a Good Thing (see Table 1). Though there is a certain contentiousness when a tester charges the vendor for allowing them access and the right to verify the samples and scenarios on which his conclusions are based.
- In fact, vendors have long shared samples with each other and with trusted testers, on the principle that the safety of the community at large takes precedence over competitive advantage (Harley, 2010b). Nor has the sharing been unidirectional,
- Some testers solicit samples/URLs from vendors. Maybe that isn't necessarily a bad thing, but it allows vendors to game a test by submitting samples other vendors are unlikely to have. Is the aim of a detection test to find out what a product *can't* detect? That doesn't sound like a bad thing from the point of view of introducing the degree of discrimination between products that sells tests, but the fact is that no product detects everything, so accurate testing needs a truly representative sample/link set that doesn't bias the results. Unfortunately, there are many, many ways to introduce bias, accidentally or otherwise: accuracy and reduction of bias were major concerns targeted by the AMTSO principles (Table 1).

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 1: AMTSO Basic Principles of Testing

Who's better at collecting and (almost more importantly) classifying and validating samples? At the level of professionalism that applies among the best-known comparative and certification testers, that's not always an easy question to answer. Collection is core functionality for both the anti-

malware and the anti-malware testing industries, and sharing is an important part of that. But it's not a full duplex process. A tester isn't likely to share/verify samples before a test. Vendors don't necessarily share samples with all (mainstream) testers, and more to the point, they don't necessarily share all samples with all other vendors in their circle of trust. So while testers tend not to have the copious resources for analysis, classification and validation that a commercial AV lab does, they may have access to a wider "common pool" than some vendors.

A sound tester is also acquainted in depth with a wide range of products, but obviously vendors (certainly on the R&D side) are pretty well acquainted with their own products: not just the mechanics behind the menu and command-lines, but also the design philosophy, the intended functionality, the dependencies between components, the implications of configuration defaults, and so on. They generally know competing products pretty well, too: obviously, they tend to do in-house comparative testing, and they have a considerable incentive to do it properly.

Discussion

AV vendors have always outnumbered testers among AMTSO members, and that was seen from the outset as "the fox in the henhouse". AMTSO has always been aware of the problem, and the Board of Directors has put a lot of effort into trying to attract more testers, but has been largely unsuccessful. One of the initial reasons for this was that outside the security industry, testing organizations tend to be concerned that *any* contact with tested vendors will compromise their neutrality, or at least be seen as doing so. Others were concerned that their ability to test effectively would be effectively neutered by having the vendors define acceptable methodology, though in fact, AMTSO guidelines documents (http://www.amtso.org/documents.html) focus on highlighting the problems with accurate testing in various areas rather than prescribing the "right" way to get round those problems.

Old Whine in New Bottles?

Before AMTSO, vendors rarely spoke publicly and in concert on poor testing. If a vendor complained publicly about a specific test, it was likely to be dismissed as vendor whining.

When vendors *did*, as occasionally happened, act in concert on testing problems (Wells et al., 2000) it was still likely to be dismissed as vendor whining: oddly, perhaps, since individual vendors wouldn't necessarily benefit in terms of product ranking from improvements in a test. In fact, companies whose products had been tested in the CNET test "under fire" in that instance and who declined to sign presumably had that likelihood in mind. Of course, the fact that they were cited in the open letter in question as having signed presumably was meant to indicate that they would have signed if there had been no such conflict of interests. The open letter is, in fact, a significant precursor of AMTSO in that included as signatories individuals involved in independent research and/or testing as well as vendors (Howes, 2001).

Nonetheless, when vendors and testers with a common interest in raising testing standards formed AMTSO, people who had always assumed that vendors always behaved with total self interest (and that testers were always beyond reproach) started to mistrust testers who actually co-operated with vendors (Townsend, 2010).

What's Been Did...

AMTSO has not been idle or totally ineffective in the past few years. Most noticeably (in concrete terms) it has generated some pretty good documentation (Table 2). The guidelines documents in the Documents and Principles repository (http://www.amtso.org/documents.html) provide community-

validated resources for testers that weren't available before, while AMTSO members have made a substantial contribution to the more general corpus of literature on the subject. AMTSO can't enforce good practice, but it's made it easier for the testing industry to conform to good practice and for the wider community to recognize what good practice actually *is*.

Document Name	Date Approved
AMTSO Fundamental Principles of Testing	31/10/2008.
AMTSO Best Practices for Dynamic Testing	31/10/2008
AMTSO Best Practices for validation of samples	7/5/2009
AMTSO Best Practices for Testing In-the-Cloud Security Products	7/5/2009
AMTSO Analysis of Reviews Process	7/5/2009
AMTSO Guidelines for testing Network Based Security Products	13/10/2009
AMTSO Issues involved in the "creation" of samples for testing	13/10/2009
AMTSO Whole Product Testing Guidelines	25/5/2010
AMTSO Performance Testing Guidelines	25/5/2010
AMTSO False Positive Testing Guidelines	22/10/2010*
AMTSO Testability Guidelines	4/5/2011

Table 2: AMTSO Documentation

Furthermore, the organization has raised general awareness of its initial concerns to such a degree that it's hard for anyone to aspire to credibility in testing while clinging exclusively to old-school static testing: indeed, there have been instances of organizations claiming to be AMTSO compliant that don't even do testing.

...And What's Been Hid

On the other hand, while most AMTSO members probably have a genuine and semi-altruistic interest in improving testing for the common good, most members also have a vested interest in some aspect of the testing process, and may be under pressure from other sectors of their organizations that are really only interested in the value to their own company. Obviously, most companies expect to get something back from their membership fee. Vendors hope that better testing will give them a better share of the positive PR that a positive test score brings, while testers hope that aligning with AMTSO's aims will demonstrate that their tests are top of the range. But over the past few years, vendors have been disappointed that particular tests haven't been "better" and in some cases have tried to use AMTSO as a lever to improve their own scores in a specific test.

The review analysis process, where tests were assessed on request for conformance with the "Principles of Testing" proved ineffective at best, alienated testers within the organization, and is currently in abeyance. So there isn't really any way in which a tester can demonstrate via an independent evaluation process that their testing is sound and accurate, and it's likely that it isn't possible to implement such a process in an organization dominated by either vendors or testers. We will consider shortly what kind of organization could evaluate, accredit or certify tests and testers more usefully: of course, there are relevant ISO standards that some AV test labs achieve, but they don't address some issues very specific to the AV and AV testing industries.

Layered Products, Layered Testing

A real problem for testers is that detection in a modern commercial product is multi-layered, and testing that only addresses one or two aspects of a product's detection technology cannot be accurate. People who read a review expect it to be authoritative, but the sad fact is that whole product testing is difficult and expensive to implement, which isn't what people who commission tests usually want to hear. For this reason testers tend to address relatively small areas of functionality in order to keep their tests manageable. A significant difficulty is in doing so without misleading the review reader into underestimating a product's abilities by artificially disabling functionality. A test audience is entitled to expect that the tester will represent accurately the functionality and value of the product or service.

A real challenge for a tester (apart from those already mentioned) is working with a test set that is truly representative of the threats that are most likely to affect the readers of its test reports, and testing in a way that accurately reflects the real world and the needs of the customer (Kaspersky, 2011). Apart from sheer sample glut of– AV labs may process hundreds of thousands of binary-unique samples a day, though the underlying code between repacked samples has not necessarily changed - there are issues like these:

- Presenting the threat in a "natural" context (one in which it's reasonable to expect a product to detect it)
- Finding a way to test detection dynamically in the cloud without risking leakage of undetected threats to external systems
- And correct classification and validation of threats and appropriate configuration of the software under threat.

No wonder that Kaspersky and (from the testing side) Myers (Myers, 2011) come to similar conclusions about the absence and unlikelihood of a truly authoritative test.

Tuning Out Static

Many magazine tests have not yet moved on from the idea that you can reliably test a product using a static test with a fixed sample set. In fact, a test with a less-than-fresh but well-validated sample set like an old-school WildList-based test may still be more useful than a poor dynamic test (Harley & Lee, 2010), but mostly in the context of meeting a proven detection baseline (product accreditation): they have little validity in terms of comparative testing, except that a product that *can't* meet what is often considered to be an "easy" test like detecting last month's WildList may be less effective than it "should" be. On the other hand, performance in any test may sometimes reflect resources poured into doing well in tests rather than improving detection performance in the field. Many people in the industry believe (Kosinár et al, 2010) that the value of comparative detection testing is quite limited because of the difficulty of testing accurately, and that it would be more

useful if testers and their audiences put more weight on other factors such as memory footprint, scan speed and so on, in order to evaluate which products might be the best fit to an individual customer's needs. But that doesn't mean that assessing performance in respects other than detection is easy! For example...

Sold a PUP

Suppose a tester includes threats that some products classify as "possibly unwanted" or a similar terminology, and leaves all the products tested at default settings. Some products don't detect PUPs/PUAs by default, preferring to leave the decision to detect them to the user (there are actually some pretty good arguments for doing this). Product A detects a PUP (say an example of adware) by default, and Product B only does so if PUP detection is specifically turned on. A test that sticks to default settings will assume that A's detection is better than B's, but in fact the test isn't comparing detection performance, but design philosophies.

Second Guessing the Bad Guys

Product A flags malware at a certain URL, but another product doesn't. Does this mean that product A is better at detection, or does it mean that the malicious server changes its behaviour when it's accessed several times from the same IP range, and stops serving malware as a defensive measure?

Quo Vadis, AMTSO?

(And is anyone going with you?)

The absence of a “voice” for the consumer and the general public may be a misunderstanding of the original aims of AMTSO, but it was one of the clubs used by “outsiders” to beat the organization with, and on occasion even by testers. AMTSO was criticized for its paternalistic mindset because of its emphasis on material generated by experts (testers or vendors). For those who regard its objectives as primarily educational, the challenge was to get useful input from (and output to) the public, the media, and even experts outside the AV and testing industries, without allowing the general level of white noise to drown the core messages.

The chosen route to meeting that challenge was to introduce a low-cost subscriber option (AMTSO, 2010) that enabled a wider range of interested parties to add their voices to the discussion. Members face a heavy burden of expectation, and the significant cost of membership – and even defaulting – is a significant incentive.

Subscribers pay less, *participate less, and less is expected from them. But the subscription model offers better communication (in theory – take-up of the opportunity has been somewhat limited to date) with the wider community, including people with a genuine interest in reading about and commenting on testing standards. All those with something to contribute to the discussion can do so by committing just enough financially to discourage the mischievous and malicious from hijacking the content creation process. Engaging with the wider community offers a far better chance of achieving AMTSO’s aims than sitting in the ivory tower telling people who aren’t listening what they’re doing wrong.

So where is the problem?

It can raise the noise level. Perhaps more importantly, some groups that have been highly influential in shaping AMTSO in the past have chosen to disengage by taking the subscription option, or dropping out altogether. Understandably, they’ve seen it as a way to disengage from the “controlling” aspects of AMTSO members within the AV industry. That may have given vendors

using AMTSO as a lever to raise their own positions in specific tests, but it can also be interpreted as demonstrating that both vendors *and* testers are sometimes less interested in raising standards than in self-promotion and self-protection.

Raising Standards

AMTSO's choice of name is a little discomfiting. It isn't, and can't be, a standards organization in the same sense that ISO is. Not, at any rate by itself. So while the "Fundamental Principles of Testing" (http://www.amtso.org/amtso---download---amtso-fundamental-principles-of-testing.html) are a decent attempt at a high-level summary of how good testing should be carried out, "compliance" is both voluntary and difficult to demonstrate. Early in the Organization's evolution it began to work on a "review of reviews" process which would make testers accountable in some sense for the accuracy and measurable "compliance" of a specific review.

Technically, however, there's no such absolute as "AMTSO compliant" and never has been, even when the organization was doing review analyses which "simply" attempted to indicate whether an individual test was in accordance with the Fundamental Principles. Such a review only measured the "compliance" (in a limited sense) of that test, rather than validating everything a tester did in some way comparable to the way ISO/IEC 17025 assesses Quality Assurance.

So there are no standards as such in process, though guidelines documents intended to address the requirements and difficulties of maintaining accuracy in various areas of security product testing have constituted a major part of AMTSO's output to date, along with a small but significant repository of external resources such as conference papers (http://www.amtso.org/related-resources.html).

More recently, there has been a great deal of interest in generating more documents aimed at helping consumers get a better idea of how to evaluate the accuracy of a test, whereas most documentation up to now has focused on helping testers improve their methodologies.

- Provision of better information
  - How (not) to test
  - Evaluating the feasibility and accuracy of a test
  - Detection technology, naming, ethics and mechanics of sample distribution
- Countering misinformation/poor test results
- Increasing the accountability of testers through certification
- Documentation: FAQs, glossaries, standards and guidelines, white papers, checklists
- Reviews of reviews
- Training and certification; standards conformance; external audit; statement of intent to comply.

There's also a lot of discussion around the organization's need to improve its own public image in order to get its messages over better to the people who need to hear them. And of course, that necessitates that the organization re-examines its own aims and frames of reference.

An AMTSO – or some other group – that *could* review tests and reviews *with credibility* could be a giant step towards meeting AMTSO's aims. But perhaps neither vendors nor testers should take the leading role in such an exercise. What if an organization (or a coalition of organizations) with more credibility (or at least a less compromised public image) were to take on the task of policing

standards enforced through certification of product certification bodies, testing organizations, and perhaps even generalist reviewers:

- To qualify for access to standard test sets?
- To prove competence, knowledge, experience and ethical fitness to test?
- To prove conformance with testing standards (AMTSO standards?), other standards e.g.
  - ISO 17024 – assessing and certifying personnel
  - ISO 9001 – quality management
  - ISO 17025 – requirements for competence of testing and calibration laboratories

In the absence of a pre-existing group with that capability, who are the stakeholders who could be part of an independent standards group? Certainly these:

- Anti-malware industry/research community
- Anti-malware testing industry
- Academia
- Standards Bodies (ISO, BSI)
- Other peripheral stakeholders in malware detection like VirusTotal and the WildList Organization.

Conclusion

It's entirely acceptable that organizations affiliated to AMTSO be accountable for attempting to conform to what the membership has approved as "good practice". But AMTSO doesn't have the muscle or, at present, the credibility to enforce standards based on that concept on external organizations, even in pursuit of generally agreed testing desiderata such as:

- Transparency and reproducibility
- Statistical accuracy based on sound metrics: sample set rightsizing, sampling techniques, metrication and instrumentation, realistic analysis, bias exclusion
- Ethical grounding: responsible disclosure, declaration of interest, sample sharing, sample generation, duty of care (safety); do no harm (do no misleading)
- Conformance to expertly formulated and agreed standards and guidelines
- Methodological validity based on comparing apples to apples rather than melons to grapes, consistency of test objectives with stated purpose, and selection of appropriate test scenarios and samples sets .
- In short, prioritization of objectivity, currency, validation, and verification of samples; reproducibility of results and methodology.

While AMTSO can and should improve its own image by positive PR, it's unlikely to become so loved that it is able to achieve testing Nirvana all on its own, at any rate while it's perceived as the AV pressure group that it has so far tried to resist becoming. But so far, it's been unable to reconcile the differing vested interests of vendors and testers within the organization (there's been bad behaviour from both sides) and some test organizations have voted with their feet. At this point it

may be best (at least in the short term) for the organization to focus on its educational role, maintaining and expanding its already impressive documentation into areas such as

Testing and security software are two sides of the same coin. But they *are* industries, and their aims are not totally compatible. The general public needs guidance on what products suit their needs best (which isn't to say that one-size-fits-all testing is necessarily good guidance).Testers need security products to evaluate, so that they can sell their results (leaving aside the unhealthily large proportion of amateur testers). (Perhaps more transparency on testing marketing models and the economic synergy between testers and vendors would benefit the consumer, though.) Vendors may not feel (or resent that) they need testers, but tests are, for better or worse, part of the marketing ecology: furthermore, *good* testing gives vendors feedback on how they're doing in terms of popularity, effectiveness etc. Actually, so does bad testing, but in that instance it's not always *useful* feedback...

In short, we are (at the top of the market) looking at two industries that know each other pretty well, and cooperate pretty well in contexts that don't encourage manipulation by intimidation and guilt-tripping. Recent discussions within AMTSO suggest that despite the partial defection of two of the biggest names in AV testing, other testers have not given up on the idea of cooperating within AMTSO, especially if AMTSO's pronouncements on tests and reviews are subject to the input of an arguably neutral third party such as academia. Even if AMTSO's effectiveness *has* been compromised by commercial interests and manipulations, a more compartmentalized model might be considered: AMTSO as an AV-dominated group working with a currently more-or-less hypothetical tester-dominated equivalent, both cooperating with other parties in an initiative under the auspices of a neutral authority with the objective of implementing true, certifiable standards where the interests of the community in general take precedence over the interests of individual vendors or testers.

*Addendum to the version published in the proceedings: it was pointed out to me in the course of the conference that some subscribers participate very actively in the forum. This is, of course, absolutely true, and the organization is the better for it. What I was intended to highlight was the fact that subscribers don't have voting rights and don't have direct access to all information exchanged internally, such as discussions on mailing lists.

References

AMTSO (2008a). Security Software Industry Takes First Steps Towards Forming Anti-Malware Testing Standards Organization. Retrieved 22 January, 2012, from http://amtso.org/amtso-formation-press-release.html.

AMTSO (2008b). AMTSO Fundamental Principles of Testing. Retrieved 22 January, 2012, from http://www.amtso.org/amtso---download---amtso-fundamental-principles-of-testing.html.

AMTSO (2010). AMTSO Widens the Conversation of Anti-Malware Testing with New Subscription Option. Retrieved 22 January, 2012, from http://www.amtso.org/pr-20101025-amtso-widens-the-conversation-of-anti-malware-testing-with-new-subscription-option.html.

Harley, D. (2010a). Antivirus Testing and AMTSO: has anything changed? 4th International Conference on Cybercrime Forensics Education and Training.

Harley, D. (2010b) Scareware and Legitimate Marketing. Retrieved 22 January, 2012, from http://blog.eset.com/2010/09/19/scareware-and-legitimate-marketing.

Harley, D. & Bridwell, L. (2011). Daze of Whine and Neuroses (But Testing Is FINE). Proceedings of the 21st Virus Bulletin International Conference PP.67-70, Virus Bulletin.

Harley, D. & Lee, A. (2008). Who Will Test The Testers? Proceedings of the 18th Virus Bulletin International Conference PP. 199-207, Virus Bulletin.

Harley, D. & Lee, A. (2010). Call of the WildList: Last Orders for WildCore-Based Testing?" Retrieved 22 January, 2012 from http://go.eset.com/us/resources/white-papers/Harley-Lee-VB2010.pdf.

Howes, E. (2001). Re: Our unique antivirus testing: How we did it. Retrieved 21 January, 2012, from http://www.dslreports.com/forum/remark,16730700.

Kaspersky, E. (2011) Benchmarking Without Weightings: Like a Burger Without a Bun. Retrieved 21 January, 2012, from http://eugene.kaspersky.com/2011/09/30/benchmarking-without-weightings-like-a-burger-without-a-bun/.

Kosinár, P., Malcho, J., Marko, R. Harley, D. (2010). AV Testing Exposed. Retrieved 22 January, 2012 from http://go.eset.com/us/resources/white-papers/Kosinar-etal-VB2010.pdf.

Myers, L. (2011). Why there's no one test to rule them all. Retrieved 21 January, 2012, from http://www.virusbtn.com/virusbulletin/archive/2011/10/vb201110-comment.

Townsend, K. (2010). Anti-Malware Testing Standards Organization: a dissenting view. Retrieved 21 January, 2012 from https://kevtownsend.wordpress.com/2010/06/27/anti-malware-testing-standards-organization-a-dissenting-view/.

Wells, J. et al (2000). Open Letter. Retrieved 21 January, 2012) from http://cybersoft.com/whitepapers/pdf/Open_Letter.pdf.