



Anti-Virus Testing Tips

Date: May 2007

Last revision: 25th May 2007

Author: Andreas Clementi

Website: <http://www.av-comparatives.org>

Test-set:

All files should be checked before they get added to the test-set. As it is (nowadays) not possible to check all files manually, ones own systems/tools should be used to check the functionality/maliciousness of the files automatically - anyway as many files as possible should still be checked (and in case of viruses replicated) manually too.

After the test the undetected samples should be sent to the tested (and trusted) anti-virus vendors, in order to give them the possibility to cross-check the results and to report known garbage files. The files reported as bad have to be taken out of the test set and RECHECKED later by hand by the testers, in order to know if they should really be taken out and also in order to prevent similar garbage from getting added in future tests (improving automated systems/tools/methods).

At AV-Comparatives we have quite limited resources and have only about 20 PC's, but we use our resources wisely, so we work much with images and removable hard disks. Our resources are compared to others limited because we did for years the tests for free (as our main goal is to inform and serve the user population), but in order to survive and further improve our work and resources we will in future also try to get paid.

Preferably do not include masses of server-side generated malware samples which you gather from malicious sites. It would be like using a virus generator to test anti-virus products. While it will increase the number of samples used for the test a lot (easily you can get a million of samples within one year by doing so), its meaning in the full test-set goes down drastically. It would probably be ok to use such samples to test the detection reliability of a specific malware family, but in a detection test it could skew the results considerably. One malware definition could bring one product to the top of the chart and another product could get penalized badly by not having yet one definition in its database, resulting in not detecting thousands of server-side generated malware samples of just one malware family.

Malware does not have an end of life time (except downloaders and some few other malwares), so a test set should contain all malware which still could cause problems to users. At AV-Comparatives we recently removed any DOS malware and malware which is definitely not causing problems anymore (like some trojan downloaders etc.) but we will still include all 32-bit malware which is a potential risk for users. Older malware is according to our tests anyway still detected at 100% by all major Anti-Virus vendors, what they do not detect is the some newer malware, which can be seen in the detection percentages.

False positive testing:

The set of clean files should be very big and include at least 10 million clean files, including also many real systems with installed software and not just files collected from CD/DVD's. The files should also be unarchived (but the archives still kept) and dupes removed. The encountered false positives should be sent to the Anti-Virus vendors (in order that they can get fixed) and get cross-checked by them, before a product gets penalized due false positives. The false alarms should be listed in a detailed report.

In some magazines we often see false alarm tests against sets of only 10000 files (chosen by the tester). Files should not be

chosen by the tester, but should include the whole set of clean files. It's a more time consuming testing but it is the only fair way to test for false alarms properly. Even more ridiculous (statistical nonsense) is it to give a false alarm rate/percentage based on a selected set of clean files.

Polymorphic virus testing:

The test set should contain at least 1000 working replicants of each virus. The amount of replicants should be equal for each virus. It should be listed which viruses were used for the test and how reliable the detection of each single virus is. Giving just a total percentage over the full set of polymorphic viruses (like we see sometimes in some magazines) is meaningless (statistical nonsense) as it does not tell if the detection of a virus is reliable or not - esp. if the amount of the used replicants varies between the used polymorphic viruses.

Methodology:

You should have and publish a methodology about how you test. The used test methodology should be available to the users together with a detailed report which explains to the users how to interpret the results. Make clear to the users that little differences in detection (like 94,7% and 96,3%) are not that meaningful (but often seen in magazine reviews), e.g. by introducing some broader detection categories in relation to the size of the used collection and taking in account the presence of some possible amount of garbage and the sources of the samples.

In any report you should state when the test was performed (missed in many printed magazines, as tests for magazines are usually at least 3-4 months old before they get printed) and which settings were used. Preferably test with default and best settings, otherwise just e.g. with best settings (to get comparable results), as e.g. some products use best settings by default and others low settings by default. You should also clearly state if you used the command line version of the product instead of the GUI version.

Detection and protection tests:

The on-demand and on-access part of an anti-virus product is often tested, but rarely we see on-execution tests. AV-Comparatives provides such kind of tests, but of course it is done only on a limited number of samples. A few vendors already participated in this kind of test, but it has not been published yet. We will take the PROTECTION (beside the detection) given by the products into account in future.

Test results should be made available for free to the users and be delivered to the public as soon as possible (of course after giving to the anti-virus vendors the time to verify the results).

Retrospective tests:

Do not make retrospective tests using only samples added to the wildlist. It will give flawed results and not reflect the real proactive detection of new malware (as the wildlist contains only some few samples and mainly variants of known malware). Retrospective tests should be made using a large number of completely new/unknown malware. It is a very complex and time consuming task to build up such a test-set but it will deliver more accurate results. Anyway, AV-Comparatives will in future (probably

2008) at least try to make retrospective tests using also e.g. 1 month old updates, to discover proactive tendencies. Some other testers will in near future start to deliver nearly real-time proactive test results and other important tests, which should be very interesting for anyone.

Outbreak tests are very important too, but they should be described in more details (like how long it took for which sample to get detected [and removed successfully]) and include a large number of malware samples.

Goals:

As tester your main goal should be to inform and help the users by providing them unbiased and independent test results of Anti-Virus software. Therefore you have not only to stay in contact with Anti-Virus vendors and other testers to possibly improve your testing methods, but also stay in contact with the users to help them understanding the results. In any case do what you can and do it as best as you can.

Finally, some NOT TO DO Tips:

- do not rely on anti-virus products to determine if a file is infected or not
- do not generate malware samples
- do not deliver test results if you have no clue about statistics and how to interpret your own results
- do not add samples to your test-set without some minimum check of the files (some few garbage may anyway slip in, but keep the rate as low as possible)
- do not put your and others PC's at risk while handling with malware (systems separated from the Internet are needed)

Copyright and Disclaimer

This publication is Copyright (c) 2007 by AV-Comparatives. Any use of the results, etc. in whole or in part, is ONLY permitted after the explicit written agreement of Andreas Clementi, prior to any publication. AV-Comparatives and its testers cannot be held liable for any damage or loss which might occur as result of, or in connection with, the use of the information provided in this paper. We take every possible care to ensure the correctness of the basic data, but a liability for the correctness of the test results cannot be taken by any representative of AV-Comparatives. We do not give any guarantee of the correctness, completeness, or suitability for a specific purpose of any of the information/content provided at any given time. No one else involved in creating, producing or delivering test results shall be liable for any indirect, special or consequential damage, or loss of profits, arising out of, or related to, the use or inability to use, the services provided by the website, test documents or any related data.

Andreas Clementi, AV-Comparatives (May 2007)