A short course in antivirus software testing: seven simple rules for evaluating tests

Sarah Gordon

Senior research fellow

Symantec Security Response

Antivirus software tests are important when selecting antivirus software. However, there are many different tests, and interpreting the results can be challenging. Additionally, the needs of the corporate customer and home user differ, and it is important to understand these differences in order to evaluate antivirus software tests critically.

Rule #1: A good test is one that is both scientific and meaningful.

Scientific tests possess several qualities. First, they have validity – they measure the thing they purport to measure. Second, they are repeatable – not only do they measure the thing they purport to measure, they do so consistently, reliably, and in ways that can be peer-reviewed. The processes are documented, and stand up to scientific scrutiny.

In addition to being scientific, a good test is meaningful. This is a bit trickier, as what is meaningful to the home user might not be meaningful to the corporate client. What is meaningful to a person from one region might not be as meaningful to a person from different region. Thus, while "meaning" is sometimes difficult to interpret the critical question is this: does the test measure something that is important to the reader? It does not matter how "in depth" a test appears to be if it is not scientific and meaningful as well. To be useful to the user, a good test must be both.

Rule #2: Not all tests are created equal.

There are several types of tests. Here we will describe several different tests, focusing mainly on good magazine tests.

University tests: University tests are an excellent opportunity for students to learn about testing processes, methodologies, and criteria by testing a wide variety of products. Results from these tests are of limited use for the corporate client or home user as they produce so much data that interpreting them correctly can be extremely time-consuming. They tend to measure things "because they can be measured". While usually scientifically valid and reliable, not all of these things will meet the "Is it meaningful to the user" criterion. Additionally, the tests are performed by students in the university environment; there is often little if any applicability of their experience or of the testing environment to the corporate world.

Commercial testers: Commercial testers offer vendors the opportunity to certify their products against criteria selected by the vendor, using a methodology approved by the vendor community and virus supplied by the vendor community (either directly or indirectly, for example via The Wild List1).

The strengths of this route are that the tests are peer reviewed, and well documented. They provide a baseline for both the corporate and home user – certifying that products detect (and when possible, repair) at a minimum the viruses that are spreading in the Wild. Additionally, when reports are available online from the commercial testing labs, such tests can be reviewed over time to show performance improvements.

Independent specialists: Independent specialists may be of value to corporate customers, but finding them can be challenging. They must possess not only intimate knowledge of antivirus software, but of the Internet, viruses and malware, and the corporate environment. They offer individualized tests and do special projects for corporate customers. Thus the output is not generally publicly available. These testers frequently go beyond the detection tests of the commercial tester and measure selected products' ability to mitigate risks specific to a particular corporation. There is a very limited number of competent independent specialists.

Magazine testers: Magazine testers are the most visible and therefore in some ways the most influential. This therefore leads us to our third rule of testing:

Rule #3: A good magazine test is subject to the same criteria as all other tests – scientific validity and meaningfulness.

Some magazine tests make use of in-house expertise to perform and interpret the tests; others hire outside contractors. In either case, there are some things to consider when evaluating the usefulness of magazine tests to a given environment.

First, the expertise of the tester should be considered. If a tester writes about modems one week, printers the next, and

antivirus the next, it is unlikely he has the expertise to test real viruses competently and safely.

As a result, magazine tests sometimes rely on the output from the commercial testers, or academic testers, and focus their own expertise on non-viral aspects of testing. This can be a useful way to approach the tests – if the test criteria and methodology meet the requirements of "scientific and meaningful".

Next, the test criteria should be evaluated for meaningfulness, and the methodology used assessed. For example, both corporate and home users need to know if a product performs against a virus they are likely to encounter. However, some "viruses" used in some magazine tests are not viruses at all; they are non-replicating or damaged samples. Measuring a product's ability to detect them has little relevance to anyone other than the tester.

Additionally, some testers use non-meaningful zoo samples, obscure or little used archivers or packers, virus simulators, or viruses created especially for a test. These all detract from the validity and reliability of the tests overall. How the criteria are chosen is extremely important.

Rule #4: A good test knows its own limits. It does not measure things just for the sake of having things to measure.

More is not necessarily better. In fact, it is usually worse. Tests that overwhelm the reader with lot of information "just because we can measure it" don't help anyone.

Rule #5: Good intentions aren't enough.

There are some things that are not well-considered by any testers. This is due to the complexity of user needs and expertise or resources required. Attempting to measure these things often result in data that are not only flawed, but misleading to both the home and corporate user. When examining test results, ensure that you pay as much attention to what is not there as to what is.

Rule #6: All things are not created equal.

It is equally important to consider how the test data are weighted in the interpretation. A product's ability to update itself automatically is important. A product's ability to detect all of the viruses in circulation is important. A product's ability to detect an obscure zoo virus sample is much less important. The ability of a product to detect a virus in an archiver is not nearly as important as its ability to detect a destructive worm entering the network.

Thus, when evaluating a test, always consider the relative weights given to different parts of the test. All things are not equally important. It is difficult to test and model all environments. Things like the System impact on detection and Synergistic/holistic effects can be very important.

Today, solutions are optimally holistic in nature. Thus, you should consider the bigger picture when evaluating any product. Would non-AV specific solutions have stopped a particular threat? Is the right response reconfiguration, firewall, or even user response?

Rule #7: Don't compare apples with oranges.

Magazine tests that measure response need to consider the needs of the reader, as well as the different types of response. Corporate and home users have very different needs and it is misleading to compare the response in one category and to make claims concerning the other. Thus it is critical that reviewers refrain from apples with oranges comparisons and do not confuse the needs of disparate groups.

A good test results in the presentation of a clear, easy to understand picture of what is being measured, and how it is being measured – and how those measurements apply to the readership's own requirements. These seven simple rules won't tell you "how" to test, or even "what" to test. However, they will help you evaluate existing tests of antivirus software critically.

In summary

Rule #1: A good test is one that is both scientific and meaningful. Does the test measure something that meaningful and is the test process from – start to finish – scientifically valid?

Rule #2: Not all testers are created equal. Does the tester have the requisite experience and knowledge to correctly evaluate the aspects of the antivirus software he is attempting to measure?

Rule #3: A good magazine test is subject to the same criteria as all other tests – scientific validity and meaningfulness – with the proviso that meaningfulness is highly contextual. Does the test accurately measure something that is meaningful to the target reader?

Rule #4: A good test knows its own limits. It does not measure things just for the sake of having things to measure. Does the test correctly interpret the data gathered?

Rule #5: Good intentions aren't enough. Does the test measure the right parts of the problem incompletely? Are the intentions good but the follow through lacking?

Rule #6: All things are not created equal. Does the test weight correctly the relative importance of the different results?

Rule #7: Don't mix apples and oranges. Does the test consistently and accurately differentiate between the home and corporate users' needs and products? Does it prevent confusion by presenting test results in light of the users' needs and product design?

About the author

Gordon is senior research fellow at Symantec Security Response. Her current research areas include testing and standards for antivirus and security software, privacy issues, cyberterrorism and psychological aspects of human/computer interaction. She will present an in-depth briefing on testing at The BlackHat Briefings in July 2004.