

# CONSIDERATIONS FOR ANTI-SPYWARE PRODUCT TESTING

## SUMMARY

This document outlines some of the considerations that should be taken into account when tests on anti-spyware products are conducted and reviewed. Tests conforming to the suggestions in this document are more likely to produce results that are scientifically valid, independently reproducible, and meaningful. It is the hope of the Anti-Spyware Coalition that this document will help testers design and perform reliable tests, and help reviewers and users interpret published test results.

## INTRODUCTION

The task of assessing whether or not a software product performs as it is designed to perform can be challenging, particularly for anti-spyware products that must protect against threats and reduce risks in an environment that is continuously changing. Software product tests can be performed for any number of reasons, ranging from comparing several products to each other, to examining just one product. Additionally, test results can be presented in a variety of ways to various audiences, from the readers of computer magazines, to the corporate security director who has commissioned a test to help her assess the product(s) that will most closely meet her corporation's needs. Finally, the tests may be performed by testers with varying levels of expertise or knowledge about this particular type of testing. Each test will have its own criteria, methodology, and purpose; yet all tests **should be scientifically valid, independently reproducible and meaningful.**

Anti-spyware software may be implemented using very different technologies. As a result, there are a range of approaches to anti-spyware product testing, from assessing simple detection of spyware to more complex risk mitigation and holistic performance. It is essential that anti-spyware product tests include one or more stated objectives that clearly describe the functions and use cases being tested. The Anti-Spyware Coalition (ASC) provides these anti-spyware testing considerations to help testers produce valid, repeatable, and useful test results.

This document can also be used to help:

- Reviewers or others who are commissioning a test to evaluate validity of the proposed test criteria, methodology and reporting.
- Consumers evaluate the relevance and usefulness of reviews or test results they read in print or online publications.
- Anti-spyware product testers evaluate the testing criteria, methodology and reporting/analysis being performed.

- Academics who are familiar with testing generically, yet have not previously been involved with anti-spyware product testing, avoid some of the pitfalls inherent in this type of testing.

## **PRE-TEST PLANNING**

In order to produce a valid, reliable test of anti-spyware products, careful planning is required. The planning process begins with defining the intended audience and objectives with as much specificity as possible.

Identification of the intended audience is essential so that users can determine the relevance of test results. Lack of clarity in this aspect of planning can result in a test that may appear relevant at first glance, but that in reality does not accurately apply to anyone. For example, “the intended audience for this test is the home user of Windows Vista 2009 in Latin America” is clear and sufficiently granular. The audience definition is essential to: develop an appropriate test methodology, select useful criteria for sample or threat inclusion, and provide context for the analysis of test results.

Testers may not know with certainty who will read their test results, but they can help users assess the applicability of the test by clearly defining, before starting the test, exactly who the intended audience is for the test. For example, if a test is designed for a small business of 50-1000 systems, the results may not apply to the home user; similarly tests designed for home users will not be as relevant for the enterprise user.

Once the audience for the test has been clearly identified, the specific objective(s) of the test should be developed as completely and concretely as possible. Defining the test objectives is a prerequisite to designing a test methodology.

Anti-spyware product test objectives may include, but are not limited to, measurement of:

- Detection of processes, files, folder, and/or Registry items
- Remediation of critical processes, files, folders, and Registry items
- Blocking/Prevention of spyware processes, installations and propagation.
- Avoidance of misidentifying non-spyware elements as spyware (false positives) during detection, remediation, and/or blocking

Test objectives should be refined to most directly address the unique needs of the identified audience. For example, a test objective to “measure and compare the ability of five consumer-oriented anti-spyware software applications to detect the top five spyware threats in Belgium today,” would aid in creating the test methodology, interpreting the results in a way that is contextually relevant and providing readers with a useful frame of reference.

Testers should refer back to these preliminary steps throughout the development of the test plan, to ensure that the test design and methodology consistently align to the audience and test purpose. Identification of the intended audience and clear definition of test goals should also be reflected in the test report to support test validity and help users determine test relevance.

## DEVELOPMENT AND DOCUMENTATION OF TEST METHODOLOGY

The validity and reliability of anti-spyware product testing is closely tied to the development and documentation of the test methodology. These general principles should guide the task of developing and documenting methodology.

1. First and foremost, the test needs sound metrics that can:
  - Measure the aspect of the product intended to be measured, providing test validity.
  - Measure the same thing consistently, providing test reliability.
  - Be performed without dependence on a personal interpretation of some aspect of a product, avoiding test subjectivity.
  - Be expressed quantitatively, in some unit of measurement using cardinal numbers.
2. Provide value in evaluating one or more aspects of an anti-spyware product.
3. If a metric does not conform to these general guidelines, the tester should reconsider the use of that particular metric in the test methodology.
4. The final analysis or interpretation of test results can be used to produce a product ranking or rating as long as the actual metrics used in the test provided quantifiable results.
5. The methodology should include choosing appropriate criteria to define the test set. For example, if the test goal is to measure a product's ability to detect spyware, the test set should not be diluted with toolbars, joke programs, or viruses. The criteria used to define the test set should always align to the test objective..
6. Testers, and organizations that publish test results, should be prepared to answer questions about their methodology of choice. Anti-spyware testers and organizations that publish test results should make information about the test methodology readily accessible.

### Choosing Test-Set Criteria

Once the test methodology and metrics are defined, a test-set of spyware needs to be obtained to test product performance. While criteria for what should be tested will differ from test to test, there are some principles that are fundamental to the appropriate selection of samples. **The test bed plays such a critical role in determining the end results that its composition should be beyond reproach.**

## Sample Selection

The identification of a test sample as an appropriate inclusion in a test set should be done independently of all products being tested; that is, no sample should be considered spyware just because another product (or products) reports it as such *nor* should it be concluded that any potential sample not detected by a product is not actually spyware. Just as an anti-virus test set should only include samples that have been independently confirmed as self-replicating, samples used in anti-spyware tests should be verified as spyware.

All of this assumes that the tester has a clear idea of what constitutes spyware. The ASC definition of spyware can be found on the ASC Web site<sup>1</sup>. As there is a broad spectrum of software that can be included within the definitions of spyware, the tester should clearly state in the test methodology the definition of spyware being used. Definitions may vary, but in the case of anti-spyware product testing, there is an easy way to tell if a program does not belong in the test set. The test set should reflect the stated goals of the test. If an object in the test set does not reflect those goals, the object should be removed. For example, a test designed to measure the ability of anti-spyware applications to detect keystroke loggers should not include simple "joke" programs or nuisance toolbars in the test set.

**The test set should only include representative samples of the types of programs defined by the test objective(s) and methodology.** The inclusion of non-representative samples can heavily skew test results. The resultant analysis of the skewed findings is unlikely to provide meaningful information about the tested product's ability to meet the stated test goals.

## Choosing the appropriate sample size

Ideally, anti-spyware testing should be conducted against a statistically significant number of samples. This should not be confused with "the more the merrier." Statistical significance is determined by a number of factors. Simply testing against "more" samples will not necessarily result in a better test. Focusing on the sheer number of samples a program deals with can result in a misleading conclusion.

More work needs to be done to determine what a statistically significant sample size is for anti-spyware testing. Until the time when there is a clear way to determine statistical significance for anti-spyware testing, each tester should explain why they believe their sample size is appropriate for the particular test.

---

<sup>1</sup> <http://www.antispywarecoalition.org/documents/documents/ASCDefinitionsWorkingReport20060622.pdf>

If the testing goals are defined appropriately, a handful of samples may be appropriate. However, if the testing goal is to assess overall effectiveness, then testing against a very small sample is not going to be statistically significant, nor appropriate for that particular test

### **Choosing samples that are prevalent and pose real threats**

Anti-spyware testing should be conducted using independently verified spyware that is currently prevalent, and poses an actual threat to real users. While there can be situations when testing against historical, isolated or non-prevalent threats may be desirable, in general the most useful tests measure threats that users are likely to encounter.

Once the test bed is assembled, testers should decide which components of the spyware programs to track and measure. Not all threat components generate the same level of concern or interest to users and administrators. Tests that treat them all equally do a disservice to the people relying on the tests, and waste the testers' valuable time. Static data files, for example, pose much less of a threat to users than executable files or dynamic link libraries. As part of the overall evaluation of a product's effectiveness, the anti-spyware tester should attempt to identify or validate whatever remnants of a malware infection remain on a test-set following the removal process. Executable code remnants are potentially far more dangerous to the end user than, for example, orphaned registry keys or files containing plain text or raw data. Evaluations should indicate the distinction clearly. Additionally, it is wise to avoid including partial or fragmentary threats in the test set. All samples should be full and complete when installed to the test environment, to ensure representative functionality, even if only a subset of critical components will be tracked and measured.

### **Potential Pitfalls**

More than a few testers have fallen prey to the pitfalls described in the following sections related to test-set validity and methodological documentation. By becoming aware of some of the common pitfalls that can lessen the value of an anti-spyware test, both readers and testers can more accurately evaluate testing criteria, methodology and results.

### **Making testing methodologies readily accessible**

Not all testers have room to publish their complete methodologies with their test results. In these cases, it is a best practice to make the complete test methodology available via an easily accessible alternative source such as a Web site. Refusal by the tester or publisher to make details about the test methodology accessible, including test criteria and any potential test biases, should cause concern for anyone reading the test results.

## Regarding simulators

Testers should avoid using artificial spyware simulation programs to test the overall effectiveness of anti-spyware products. The use of simulators can lead testers to inaccurate findings because many anti-spyware products are designed to detect behaviors that are difficult to encapsulate in a simulated environment. Using simulators (including virus simulators, Trojan simulators, "leak" testing programs, and behavioral simulators) can unfairly penalize those programs whose design does not easily fit the underlying assumptions behind the simulators. The ability of a product to detect fake or simulated threats is not useful in the real world. Use of simulators could incentivize developers to build programs that satisfy the test rather than combating real world malware.

## CONFIGURATION OF PRODUCTS BEING TESTED

Although anti-spyware applications tend to have a similar range of functionality and options, applications can vary in how that functionality is implemented. Moreover, some anti-spyware applications may have features not found in the other applications selected for testing. These may or may not be relevant to the test. There are two approaches to managing differences among applications:

- The first approach is to test programs of similar functionality and technological implementation. In this scenario the tester may configure the applications as similarly as possible or may choose to use the product's default configurations. Commercial certification labs (eg. ICSA Labs, Virus Bulletin, West Coast Labs) tend to favor default configuration testing as it is indicative of how many users will implement the products and can provide a valid baseline test.
- The second approach is to test using a more holistic test design. The applications are configured to give the best possible detection and remediation performance in a given situation, and any aspects of the program that contributes to the program's effectiveness are tested. This type of test provides the most complete and useful results, but may be impractical due to cost, and level of expertise required.

Some product features may not always be appropriate to a test's goals. For example, a test designed to help home PC users select the most effective and "easy-to-use" anti-spyware application should probably not use the applications' heuristic detection abilities if they require the user to make frequent decisions that fall beyond the scope of the target audience's technical knowledge.

Testers should also consider their approach toward security software application suites, which are already quite common in the consumer anti-malware market and are becoming more prevalent in the small- and medium-sized business and enterprise

markets. Optimally, if an application suite is being tested, the test should consider performance holistically as this type of testing provides a more accurate picture of how the tested product will actually perform when used as designed. Simply put, measurement of one aspect of a multi-faceted product is an insufficient metric in determining to what degree any product, when correctly implemented, will contribute to the reduction of the users overall risk in a real-world scenario.

Whenever there are issues with differences in features and functionality among the tested applications or suites, decisions as to how to manage those differences should be guided by the test objectives and target audience.

Finally, any test of anti-spyware products needs to include some way to measure false positives. False positives can have significant impact on customers. For example, they can inappropriately remove a critical system file thereby disabling a system, or remove a component of a desired application that renders it inoperable. Without false positive testing reflected in the methodology and test criteria, a product could raise an alert on every downloaded or running program in the test and still get a perfect score. If a test deems a non-zero level of false positives acceptable, the tester should describe the assumptions by which this conclusion was reached for each product tested where a false positive was present. Once again, putting results in context is important.

## TEST ENVIRONMENT

Generally speaking there are two types of test environments used in anti-spyware product testing:

- **Static test environments** in which spyware samples are simply placed on a disk for use in scan testing and the samples are not executed, and
- **Dynamic test environments** in which spyware samples are executed or installed live.<sup>2</sup>

Static test environments are far easier to set up, maintain and use, and more easily provide reproducible results. Dynamic environments more closely approximate what users and administrators will actually encounter when dealing with spyware.

Most important is that the test environment used should resemble the environment of the intended audience(s) as closely as possible. The environment should consist of hardware that approximates what the intended audience actually uses, and should be installed with operating systems and user applications that would be typical.

---

<sup>2</sup> In dynamic tests, the sample should be verified as active, and not just present on the computer (i.e., if it is a keylogger, it should actually be logging keys; if it is a password stealer, it should actually be stealing passwords).

Further, whatever test environment is used, it should be constructed so that it lends itself to multiple test runs with minimal variation. Some spyware is delivered on a rotational basis from servers that generate randomly unique variants on the fly, greatly complicating the task of testing anti-spyware products within a dynamic test environment. In such cases, testers may need to use controls to standardize testing of samples.

Regardless of the test environment selected or created, a small, well-chosen test set of spyware tested in a dynamic environment will provide more useful and meaningful results than a large, unverified test set scanned in a static environment. (See earlier sections about test samples for additional information.)

### **Virtual Machine Environments**

Performing testing in virtual machine environments, such as those offered by VMware or Microsoft Virtual Server, offers convenience to the tester, but may affect the accuracy of results. On the positive side, virtual machines enable testing with limited resources and make it easy to ensure a consistent test environment. However, some spyware samples have been shown to behave differently in virtual computing environments than they do in native environments.

Given that more spyware is designed to detect virtual environments and refuse to execute on such configurations, and given that many malware programs use advanced features, such as stealthing technologies, that may not function properly in virtual environments, it may be necessary (depending on the test set) to employ real physical machines that are re-imaged between tests. In doing so, it is important to ensure that machines used do not already contain malware that is not part of the actual test-set.

If testers wish to use virtual environments, they should perform initial testing to compare their spyware samples' behavior within both types of environments. They should also disclose their use of virtual environments in their test results. An alternative that offers some of the convenience of virtual environments, but without the risk of altered results, is to use hard drive images.

## **TEST ANALYSES AND REPORTING**

The testing process is not complete until the finished report is available for review or challenge.

Key aspects of any of anti-spyware test report are transparency, readability, and validity. The report should detail the testing process and methodology in a clear and comprehensible manner. In a summary form, the test report should accurately reflect testing logs and other documentation that, if one or more test results are called into question, support the findings published in the report.

**Researchers, testers and reviewers should provide a readable, useful report that explicitly describes the key attributes of the test, including (but not limited to):**

- intended audience
- test goals/objectives
- design/methodology
- criteria for threat/risk inclusion
- criteria for anti-spyware application inclusion
- metrics
- test execution
- method of data analysis.

Additional principles that should be followed when producing the test report include:

- The resultant raw data should be interpreted in the context of the stated goals of the test.
- The report should disclose any potential test bias, or areas of potential dispute or conflicts of interest.
- Limitations of the test design should be clearly described.
- The report should identify and correctly cite all external sources & authorities used, drawn upon, quoted, or even relevant to the test design, execution, and report.

By anticipating, acknowledging, and responding to potential objections and criticisms of the test design and execution, testing can be improved, ultimately benefiting both the tester and the user.

## **CONCLUSION**

Anti-Spyware testing must produce scientifically valid, independently reproducible, and meaningful results. The test must be carefully designed in order to answer specific questions. The test methodology must be carefully developed and documented in order to make the test understandable to users, reviewers, and testers. The objective of the test, as well as the methodology used, must be clearly explained. By placing the Anti-Spyware product test results in the context of the objective of the test and the testing methodology, users, reviewers, and testers will be able to interpret the test results more effectively. Through careful decisions, testers will produce more reliable and valid test results.