Microsoft

Transparency report

Examining the AV-TEST March-April 2018 results

Prepared by

Windows Defender Research team

Microsoft

Table of Contents

Microsoft

1 Introduction

This report presents Windows Defender Antivirus (Windows Defender AV) test scores in AV-TEST's March-April 2018 testing cycle with commentary for context and transparency.

1.1 Key takeaways

Below is a summary of this report:

Protection: Windows Defender AV achieved an overall Protection score of 5.5/6.0, missing 2 out of 5,680 malware samples (0.035% miss rate). With the latest results, Windows Defender AV has achieved 100% on 9 of the 12 most recent tests (combined "Real World" and "Prevalent malware"). Learn More

Usability (false positives): Windows Defender AV maintained its previous score of 5.5/6.0. Based on telemetry, most samples that Windows Defender AV incorrectly classified as malware (false positive) had very low prevalence and are not commonly used in business context. This means that it is unlikely for these false positives to affect enterprise customers. Learn More

Performance: Windows Defender AV maintained its previous score of 5.5/6.0 and continued to outperform the industry in most areas. These results reflect the investments we put in optimizing Windows Defender AV performance for high-frequency actions. Learn More

2 Dissecting test results

2.1 Summary of overall scores

The table below summarizes Windows Defender AV's overall test results in the March-April 2018 testing by AV-TEST:

	Protection	Usability	Performance
Overall score for this cycle >>>	5.5/6.0 (-0.5)	5.5/6.0 (±0)	5.5/6.0 (±0)

Table 1. Windows Defender AV's overall test results in the March-April 2018 AV-TEST Business User test. AV-TEST uses Protection, and Usability, and Performance test modules.

2.2 Understanding Protection scores

Below are details of Protection scores in the March-April 2018 testing cycle:

	March	April
"Real World" testing	98% (100/102)	100% (94/94)
"Prevalent malware" testing	100% (2,495/2,495)	100% (2,989/2,989)
Overall malware protection rate (all samples)	99.96% (5,678/5,680)	
Overall Protection score for this cycle >>>	**5.5/6.0**	
Overall Protection ranking for this cycle >>	11th out of 15 (tied with 3 others)	

Table 2. Summary of Protection scores for the March-April 2018 Business User test

The diagrams below show Windows Defender AV's detection rates in "Real World" and "Prevalent malware" testing over a one-year period:

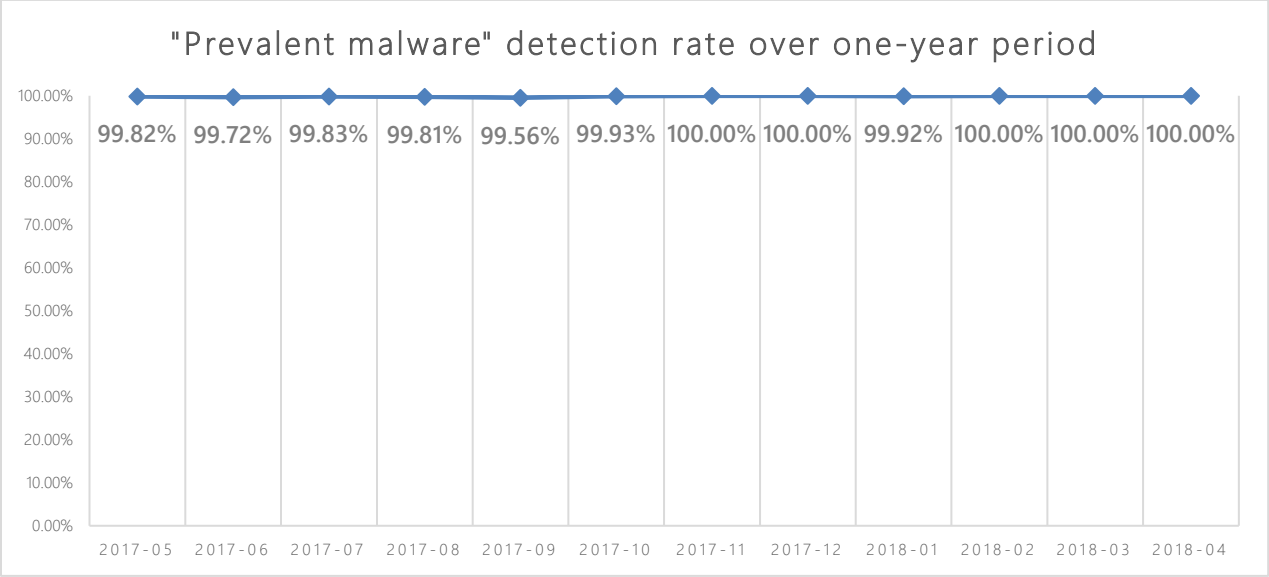


Figure 1. Windows Defender AV detection rates in AV-TEST "Prevalent malware" tests over a one-year period

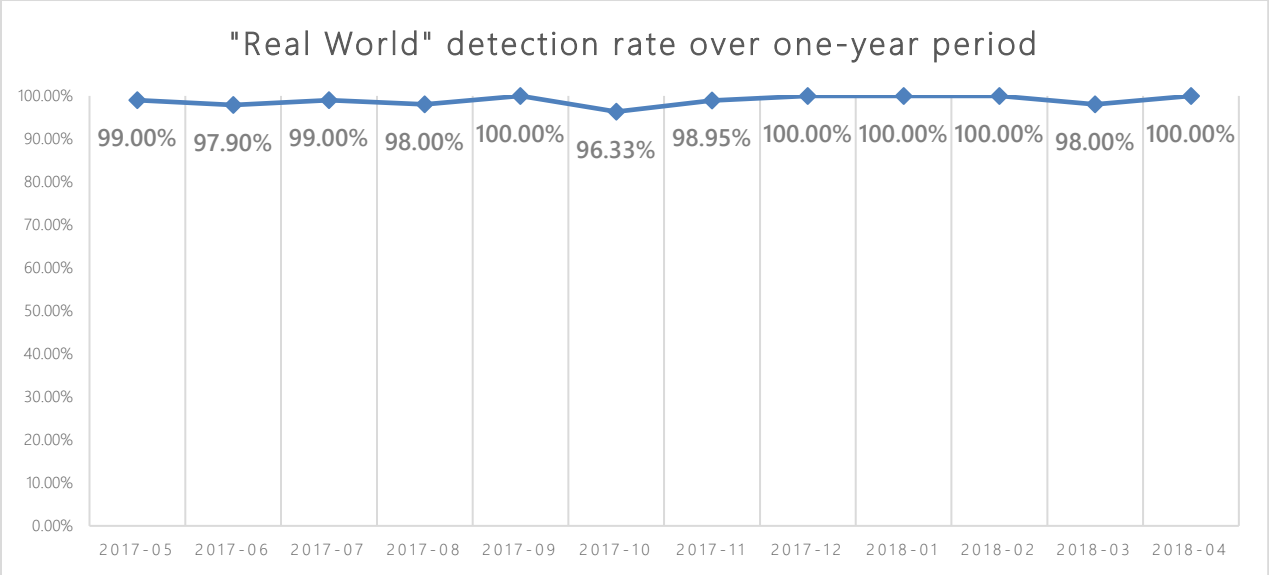


Figure 2. Windows Defender AV detection rates in AV-TEST "Real World" tests over a one-year period

2.2.1 Missed samples are opportunities for improvement

Windows Defender AV missed 2 out of 5,680 samples used in the Protection test module. The Windows Defender Research team takes missed samples as an opportunity to improve detection capabilities. For each missed sample, a team of researchers analyzes and assigns a correct verdict to the sample to make sure it is detected. In addition, the team also analyzes the root cause for the miss and drives long-term detection improvements.

Below is the analysis of the missed samples and the improvements made as a result:

Missed Sample	Miss reason	Improvements
Sample 1	Even though this sample was detected by Windows Defender AV, a bug that occurred only in very specific conditions led to incomplete remediation of the sample. The bug was related to prioritization of concurrent signature matches	• The bug has been fixed since in an engine release in May. Full remediation now takes place under similar conditions
Sample 2	No malicious behavior match on the sample	• New behavioral triggers added • JavaScript Antimalware Scan Interface (JAMSI) detections improved

Table 3. Improvements made to Windows Defender AV as result of this cycle's test results

2.2.2 Examining "Real World" tests

There are important factors to consider when interpreting "Real World" test scores:

- **The size factor:** Windows Defender AV encounters a staggering ~200 million samples every month, 96% of which are polymorphic. The vastness of the malware landscape makes it extremely difficult to evaluate the quality of protection against real world threats, especially given that typical sample sets consist of ~100 samples.

Isolating AV from the rest of the Windows Defender ATP stack discounts the synergy among components and creates conditions that don't reflect the real world

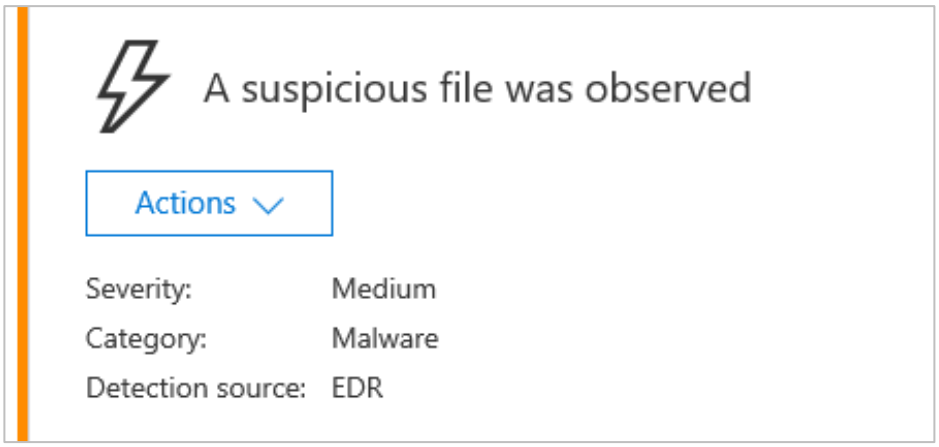
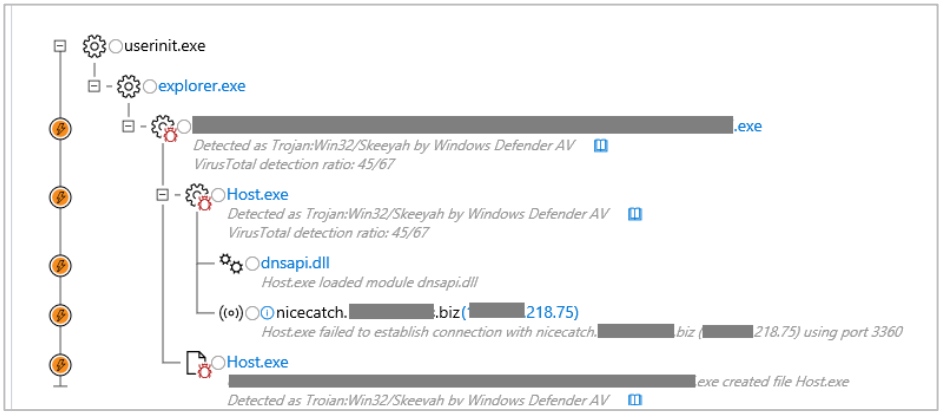
- **The synthetic conditions factor**: Synthetically emulating real-world infection often discounts contextual and behavioral clues that normally accompany malware infections in the real world. Such clues are important for behavioral detections and their absence can reduce the effectiveness of security solutions.

- **The isolation factor:** Isolating AV also discounts the synergy with other protection components in real enterprise networks. The Windows Defender Advanced Threat Protection (Windows Defender ATP) stack includes components like endpoint detection and response (EDR) capabilities, Windows Defender SmartScreen, Windows Defender Exploit Guard, and others.

2.2.3 True real-world testing with the Windows Defender ATP stack

The Windows Defender AV team tested the two missed samples against the Windows Defender ATP stack to assess the samples' ability to infect machines in real-world enterprise environments. This expands on the testing practice that isolates AV from the rest of the environment. As expected, the malware samples were blocked and detected by several stack components, as follows:

Sample 1:

Windows Defender ATP component	Test outcome
Endpoint and detection response (EDR) capabilities in Windows Defender ATP (Windows Defender Security Center)	Alert generated: "A suspicious file was observed"  Alert process tree shows infection chain: 

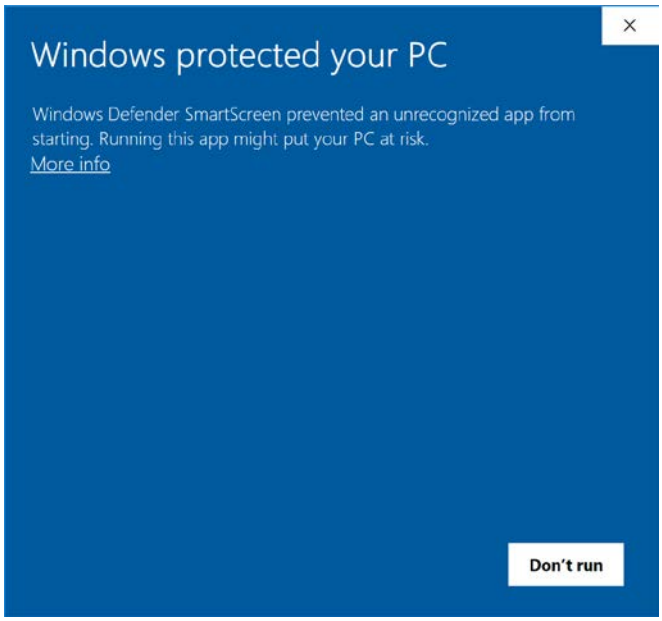
Windows Defender SmartScreen	File blocked from running using the AppRep capability: 
Windows Defender Application Control	File blocked from running under the following modes: - Intelligent Security Graph Mode - Whitelisting mode - Managed Installer mode
Windows Defender Application Guard	File blocked from being downloaded and run from the web

Table 4. Running sample 1 against the Windows Defender ATP stack

Sample 2:

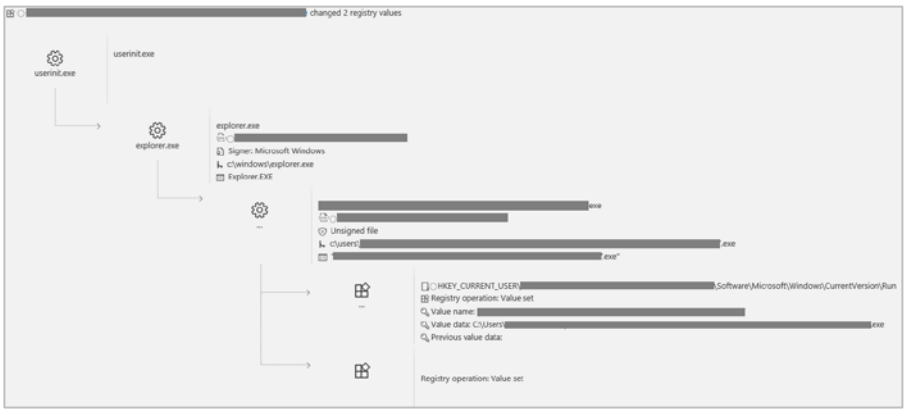
Windows Defender ATP component	Test outcome
Endpoint and detection response (EDR) capabilities in Windows Defender ATP (Windows Defender Security Center)	Alert Generated: "Suspicious behavior by a scripting tool was observed" Alert process tree shows infection chain: 
Windows Defender Application Control	File blocked from running in the following modes: - Whitelisting mode - Managed Installer mode
Windows Defender Application Guard	File blocked from being downloaded and run from the web

Table 5. Running sample 2 against the Windows Defender ATP stack

2.3 Understanding Usability scores

In Usability tests, AV-TEST includes clean file samples in the test population and checks whether antivirus products incorrectly classify them as malware (what is known as false positive, or FP). Below is a summary of Windows Defender AV results in the Usability test:

	March	April
Number of misclassified files	5 (out of 402,861 samples)	2 (out of 417,596 samples)
Overall score for this cycle >>>	5.5/6.0 (±0) ←→	
Overall Ranking >>>	10th out of 15	

Table 6. Usability test results summary for Windows Defender AV for the March-April cycle

2.3.1 Analysis: What kinds of files did Windows Defender AV misclassify?

Of the seven clean file samples that Windows Defender AV incorrectly classified as malware, three were not observed in any Windows Defender AV customer in March or April. Furthermore, only two of these files were blocked in actual enterprise environments, affecting less than 12 machines in total. Overall, based on our research and the file prevalence data, most of the samples that Windows Defender AV misclassified are not common in enterprise environments. Below is a sample list of files that Windows Defender AV misclassified in this test cycle.

Sample	File prevalence (30 days)	Description
Sample a	2	Billing tool for midwives (unsigned)
Sample b	0	Highlighter tool for Excel (unsigned)
Sample c	3	Outlook add-in for inserting frequently used text in emails (unsigned)
Sample d	3	Codecs package (unsigned)

Table 7. Files that Windows Defender AV incorrectly classified as malware

Microsoft encourages software vendors to sign their software with certificates issued by reputable Certification Authorities. This will raise the level of trust both by security vendors and users alike.

2.3.2 The synthetic nature of Usability tests

Misclassifications in a synthetic test are not necessarily indicative of false positives in real-world scenarios. This is true when the test methodology discounts contextual elements that Windows Defender AV uses for issuing a verdict. For example, when a file is tested, it is not downloaded from the vendor website. Both the original file name and the download site are contextual information that are removed in tests. We've seen many cases where a customer in the real world downloads a clean program from the vendor site without encountering any erroneous detection. However, when a tester gives the file a seemingly random name (e.g., it's SHA-256), removes the mark of the web, and doesn't download the file from the vendor website, some of our more aggressive machine learning models issue blocks that don't occur in the real-world.

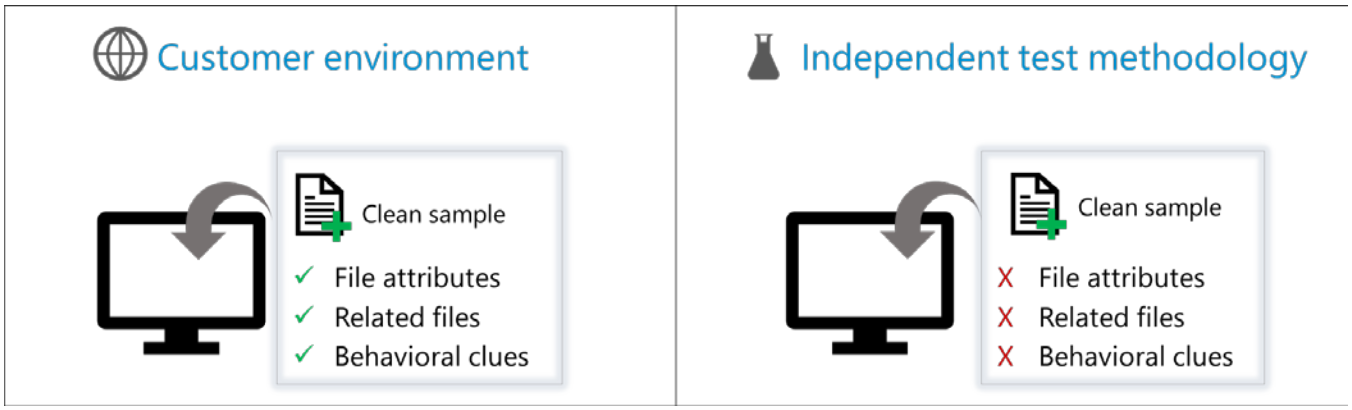


Figure 3. In some cases, Windows Defender AV incorrectly classified samples (false positive) in the synthetic test environment but not on customer machines.

2.3.3 Criteria for evaluating files may vary across vendors and testers

The criteria for classification can vary between antivirus vendors and testers depending on their policies. Some of the files identified as clean by some vendors could be files that Windows Defender AV identifies as potentially unwanted application (PUA) and thus would be blocked. Microsoft's policy aims to protect customers against malicious software while minimizing the restrictions on developers. The diagram below demonstrates the high-level evaluation criteria Microsoft uses for classifying samples:

- Malware: Performs malicious actions on a computer.
- Unwanted software: Exhibits the behavior of adware, browser modifier, misleading, monitoring tool, or software bundler
- Potentially unwanted application (PUA): Exhibits behaviors that degrade the Windows experience
- Clean: We trust that the file is not malicious, is not inappropriate for an enterprise environment, and does not degrade the Windows experience

Figure 4. Microsoft's high-level sample classification criteria

2.4 Understanding Performance scores

Performance tests measure the effect of certain user actions, which are executed as part of the test, on system speed. The table below summarizes Windows Defender AV's Performance test results in the March-April cycle:

	March-April
Overall Performance Test Score	5.5/6.0 (±0) ←→
Product Ranking	8/15 (tied with 4 more vendors)

Table 8. Performance test results for Windows Defender AV for the March-April cycle

The table below presents the details of performance test results compared to industry averages. Performance is measured by the average impact of the product on computer speed. Therefore, a smaller number is favorable. Green boxes indicate areas where Windows Defender AV performed better than the industry average; red boxes indicate where Windows Defender AV performed lower than the industry average.

Action	Standard PC	Industry Average	High End PC	Industry Average
Launching popular websites	5%	14%	4%	13%
Downloading frequently used applications*	0%	0%	0%	0%
Launching standard software applications	12%	11%	15%	11%
Installation of frequently used applications	51%	32%	43%	32%
Copying of files (locally and in a network)	3%	6%	2%	7%

Table 9. Average impact of the product on computer speed in daily usage

*The description for these operations is given by AV-TEST and might not be aligned with what Microsoft's data indicates as realistic.

2.4.1 Areas that matter most to customers

Based on results presented in Table , Windows Defender AV outperformed the industry average in most areas. The only area where Windows Defender AV performance is substantially below the industry average is in *Installation of frequently used applications*.

There are several factors to consider for driving the right conclusion out of these test results:

- **Consider the frequency of the action**
  Most users in enterprise environments are information workers whose common user activities include:
  - Browsing the web
  - Using email clients
  - Processing documents
  - Accessing network resources

  Users spend substantially less time installing new applications compared with the activities listed above. This is true for all user segments, but it is especially true for enterprise users where software installation is usually governed by usage policies. Windows Defender AV's performance is optimized for delivering high levels of performance in high frequency actions for better overall user experience. This is evident in the data presented in Table 9, where *Installation of frequently used applications* (a low-frequency action) is the only area where Windows Defender AV scored substantially lower than the industry average. In *Launching standard software applications*, despite scoring slightly lower than industry average, overall impact has decreased since the previous cycle, reflecting improved performance by Windows Defender AV. Performance is a priority area for the Windows Defender AV team, and we're working to improve it even further.

- **Consider the level of risk**
  Windows Defender AV is designed to perform thorough scanning during the software installation process. This could have a performance cost, as shown in Table 9. One reason for this is that software installation is a relatively complex operation that touches different areas of the operating system. Thorough inspection is necessary to address the risk of introducing malicious software on the system.

- **What impactful areas are not being tested?**
  There are several areas that are not being tested for performance by AV-TEST that are critical to user experience. Examples include:

  - Shutdown and startup
  - Universal Windows app launch
  - Battery consumption