ARE THEY REAL?

REAL-LIFE COMPARATIVE TESTS OF ANTI-VIRUS PRODUCTS

Fanny Lalonde Lévesque, École Polytechnique de Montréal

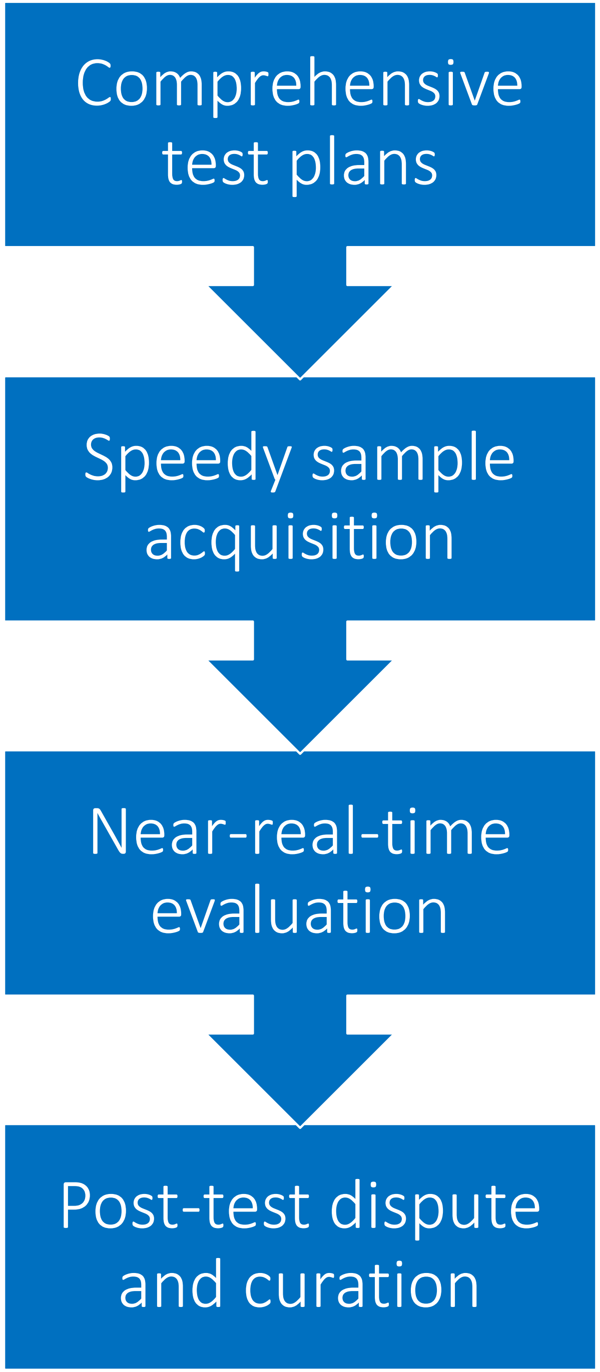
José Fernandez, École Polytechnique de Montréal

Glaucia Young, Microsoft

Dennis Batchelder, AppEsteem Corporation

Virus Bulletin Conference
5 October 2016

Introduction



Comprehensive test plans

Speedy sample acquisition

Near-real-time evaluation

Post-test dispute and curation

The AV comparative testing industry has developed best practices for squeezing the most out of their lab tests

But it’s not enough…

It’s hard to account for what happens to real people in real life

Measuring real-life usage would allow us to:

Understand

- Rank AV effectiveness across different customer segments
- Measure human/environmental impact on results
- Compare AV effectiveness in the lab vs real-life
- Measure how product design features impact effectiveness

Answer some hard questions

- Are AVs a commodity or not?
- Does paid vs. free AV make a difference?
- Will a monoculture help the bad guys?

Unfortunately, real-life approaches like clinical studies are too small in scope and take too long to get results.

So we constructed a study that used Microsoft’s telemetry to conduct a large-scale, real-life AV comparative test.

Test goal: Measure AV effectiveness against well-known malware (hygiene) on real customer systems

Study design and methods

World's first large-scale "real-life comparative evaluation"

Study design

Study population

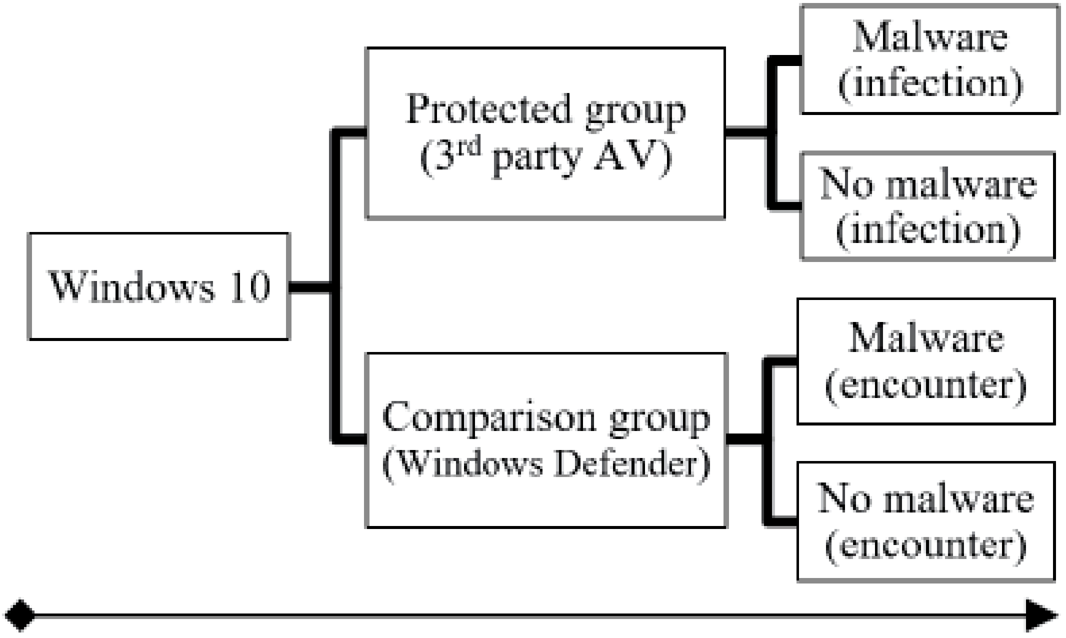
- Windows 10 systems

Protected group

- Systems protected by a 3rd party AV
- Outcome: malware infection

Comparison group

- Systems protected by Windows Defender
- Outcome: malware encounter (proxy for no AV)



4 months

Real-life data collection

Study population

- 26M Windows 10 systems
- November 2015 – February 2016

Protected group

- 16M systems protected by a 3rd party AV
- Outcome: Based on MSRT infections

Comparison group

- 10M systems protected by Windows Defender
- Outcome: Based on Windows Defender encounters

Machines included in the study

- Single user machines only
- PCs, laptops, tablets (no phones/XBox)
- Non-shared machine GUIDs
- Known age group and gender (based on Microsoft Account demographics)
- Only countries that have HDI data
- Observed for the entire test period
- Kept the same AV for the test period

Malware families included

- Malicious and unwanted software
- Covered by MSRT for at least one month prior
- Threats with known categories

Study population

Factors selection

- Relatable: readers of the test can self-select
- Simple: don’t over-slice
- Durable: don’t change these often, so we can construct a history for trending

User factors

- Gender (2)
- Age group (5)

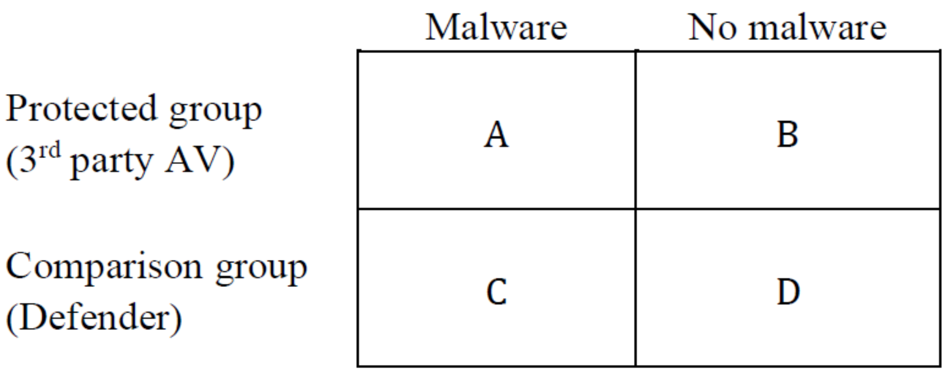
Environmental factors

- Region of the world (6)
- Country’s United Nation’s Human Development Index (4)

Factors	Protected group	Comparison group
Gender		
Female	35.90%	35.02%
Male	64.10%	64.98%
Age group		
0-17	4.57%	5.73%
18-24	18.16%	21.04%
25-34	20.70%	24.24%
35-49	25.55%	25.04%
50+	31.02%	23.95%
Region		
Africa & Middle East	1.76%	2.77%
Asia & Pacific	11.62%	11.24%
Australia	2.54%	2.31%
South & Central America	7.55%	6.95%
North America	39.54%	43.73%
Europe	36.99%	33.00%
HDI category		
Very high	81.63%	79.73%
High	15.98%	15.69%
Medium	2.14%	3.95%
Low	0.24%	0.63%

Calculating anti-virus effectiveness (AVE)

Step 1 : Frequency of malware infection



	Malware	No malware
Protected group (3rd party AV)	A	B
Comparison group (Defender)	C	D

Step 2 : Relative risk of malware infection

$$RR = \frac{{}^{A}/_{(A+B)}}{{}^{C}/_{(C+D)}}$$

Step 3 : Anti-virus effectiveness

$$\text{Effectiveness} = (1 - RR) \text{ x } 100$$

A : number of systems in the protected group that got infected by malware

B : number of systems in the protected group that did not get infected by malware

C : number of systems in the comparison group that encountered malware

D : number of systems in the comparison group that did not encounter malware

Results

Anti-virus effectiveness primary analysis

26,956,360 unique systems were assessed over 4 months

Protected group

- 16,464,720 systems
- 201,517 systems got infected by malware (1.22%)

Comparison group

- 10,491,630 systems
- 1,568,122 systems encountered malware (14.85%)

Estimated effectiveness of all 3rd party AVs*

91.81%

$$\mathrm{RR} = \frac{201{,}517/16{,}464{,}730}{1{,}568{,}122/10{,}491{,}630} = 0.0819$$

$$\mathrm{AVE} = (1 - 0.0819) \text{ x } 100 = 91.81\%$$

*Windows Defender AVE can’t be calculated with this method

Anti-virus effectiveness by factor

AVE differs by factor (see table)

- AVs much more effective for malicious software
- AVs more effective for males
- AVs most effective for 25-34 and least effective for 0-17
- AVs most effective in Asia and least effective in North America

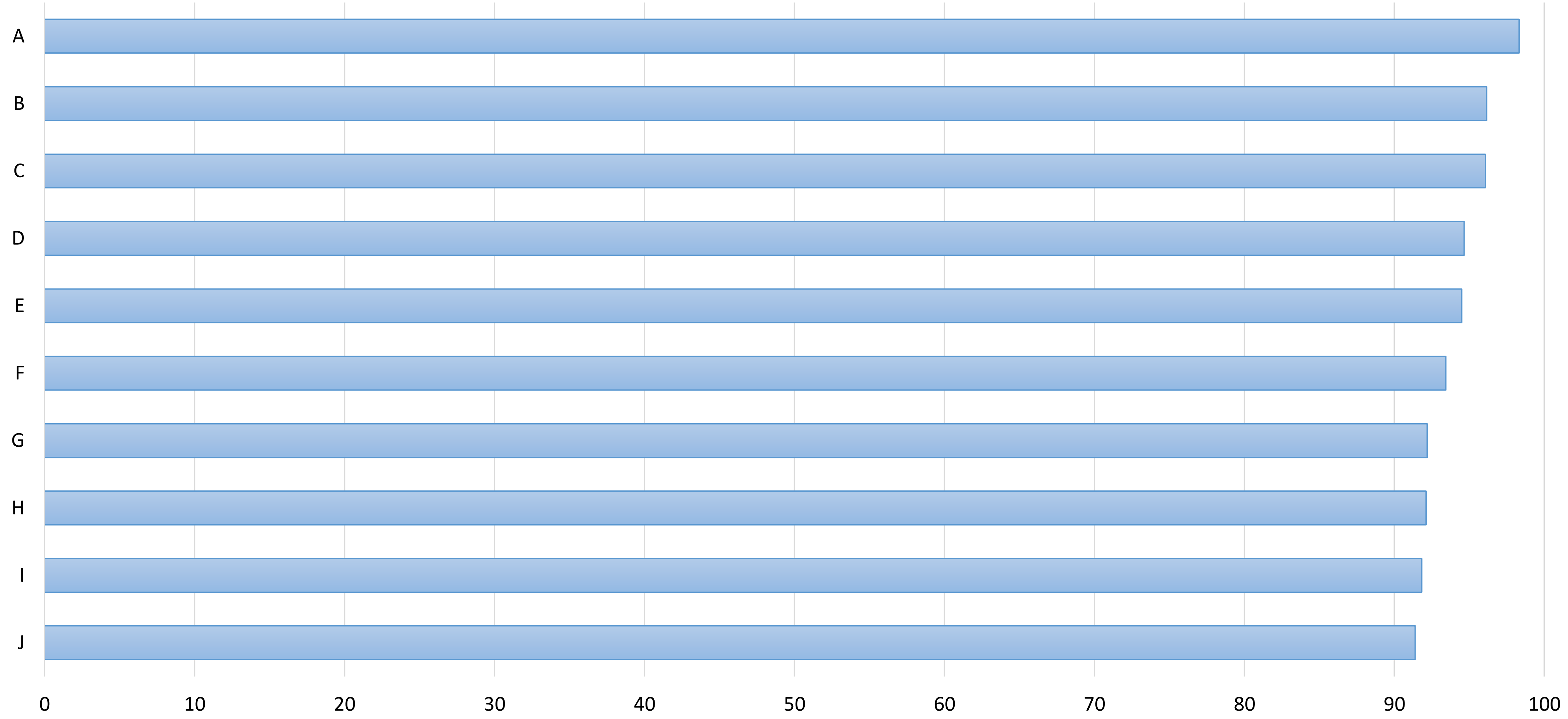
Combining factors together yields more understanding

- 50+ more infected with rogue malware and ransomware

Factors	AVE
AV protection status	
Full	91.93%
Partial	89.80%
Malware types	
Malicious software	99.47%
Unwanted software	56.39%
Gender	
Female	89.39%
Male	92.54%
Age group	
0-17	87.65%
18-24	91.94%
25-34	92.27%
35-49	91.25%
50+	90.80%
Region	
Africa & Middle East	92.09%
Asia & Pacific	96.17%
Australia	88.52%
South & Central America	93.29%
North America	87.91%
Europe	91.76%
HDI category	
Very high	88.72%
High	95.44%
Medium	92.64%
Low	94.51%

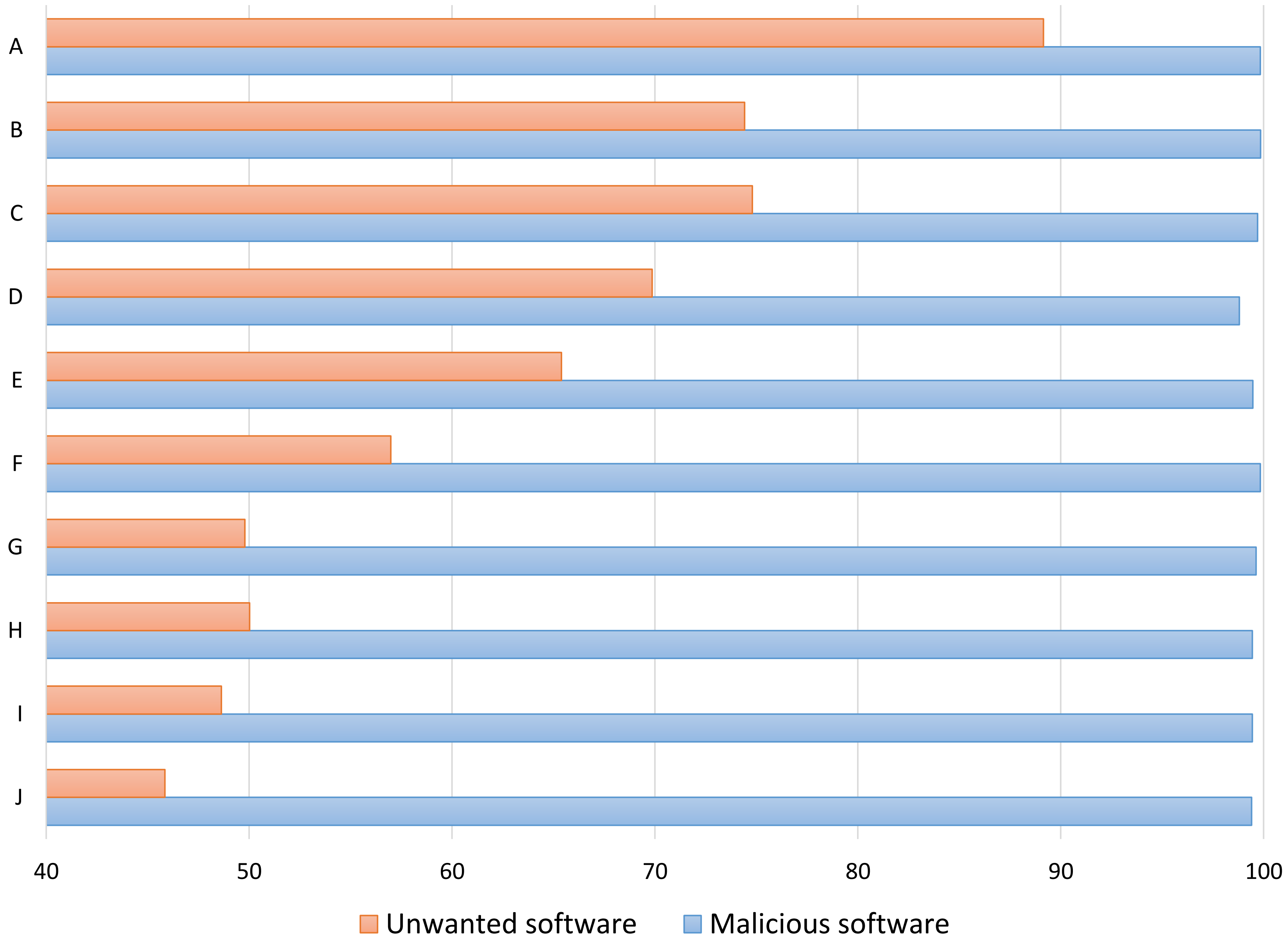
Anti-virus effectiveness comparative analysis

10 most prevalent 3rd party AVs



Anti-virus effectiveness by malware type

- Similar AVE for malicious software
- Important variations in AVE for unwanted software
- Vendors who performed better for malicious software also performed better for unwanted software

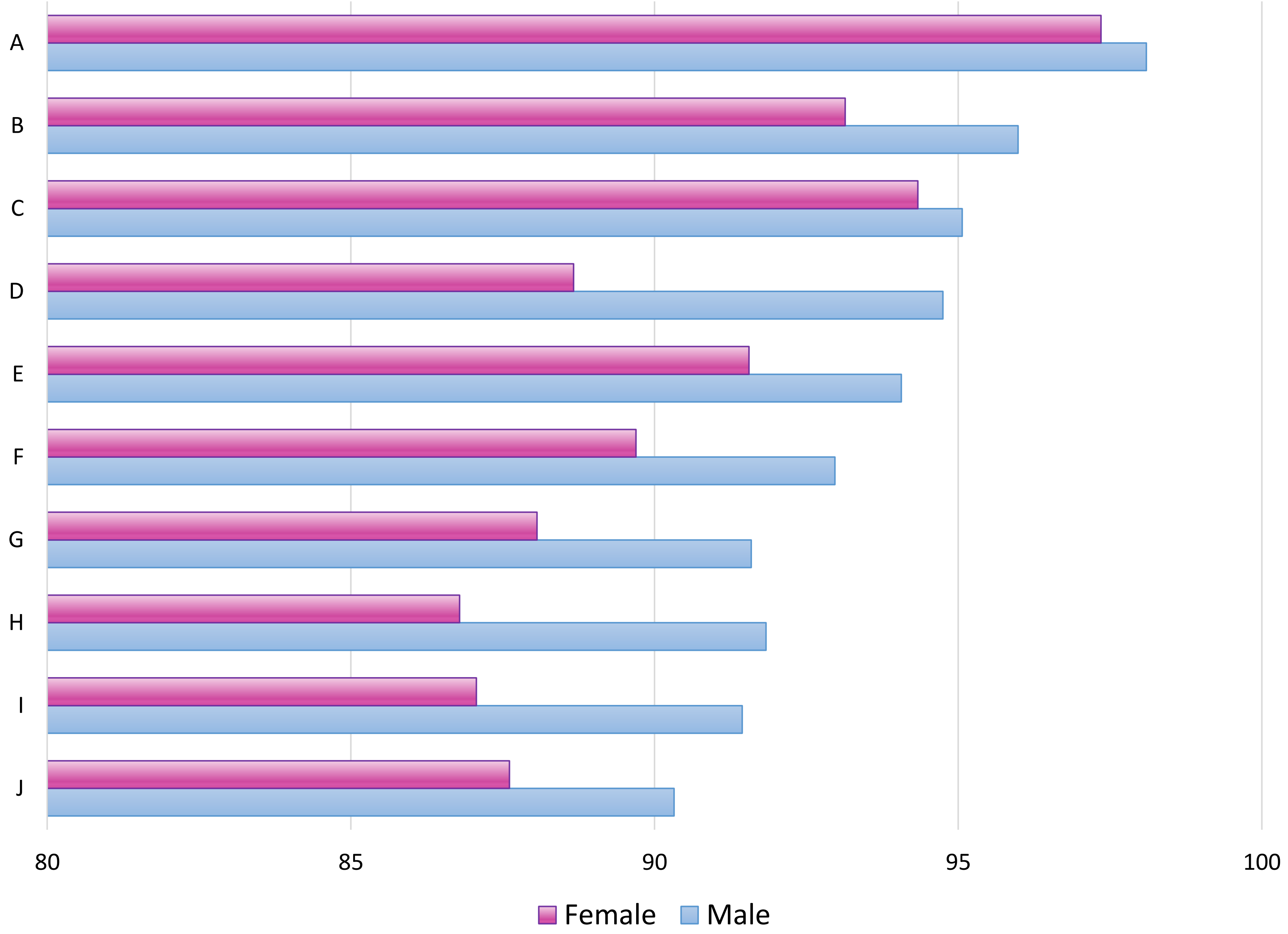


Anti-virus effectiveness by gender

- Every 3rd party AV was more effective protecting males

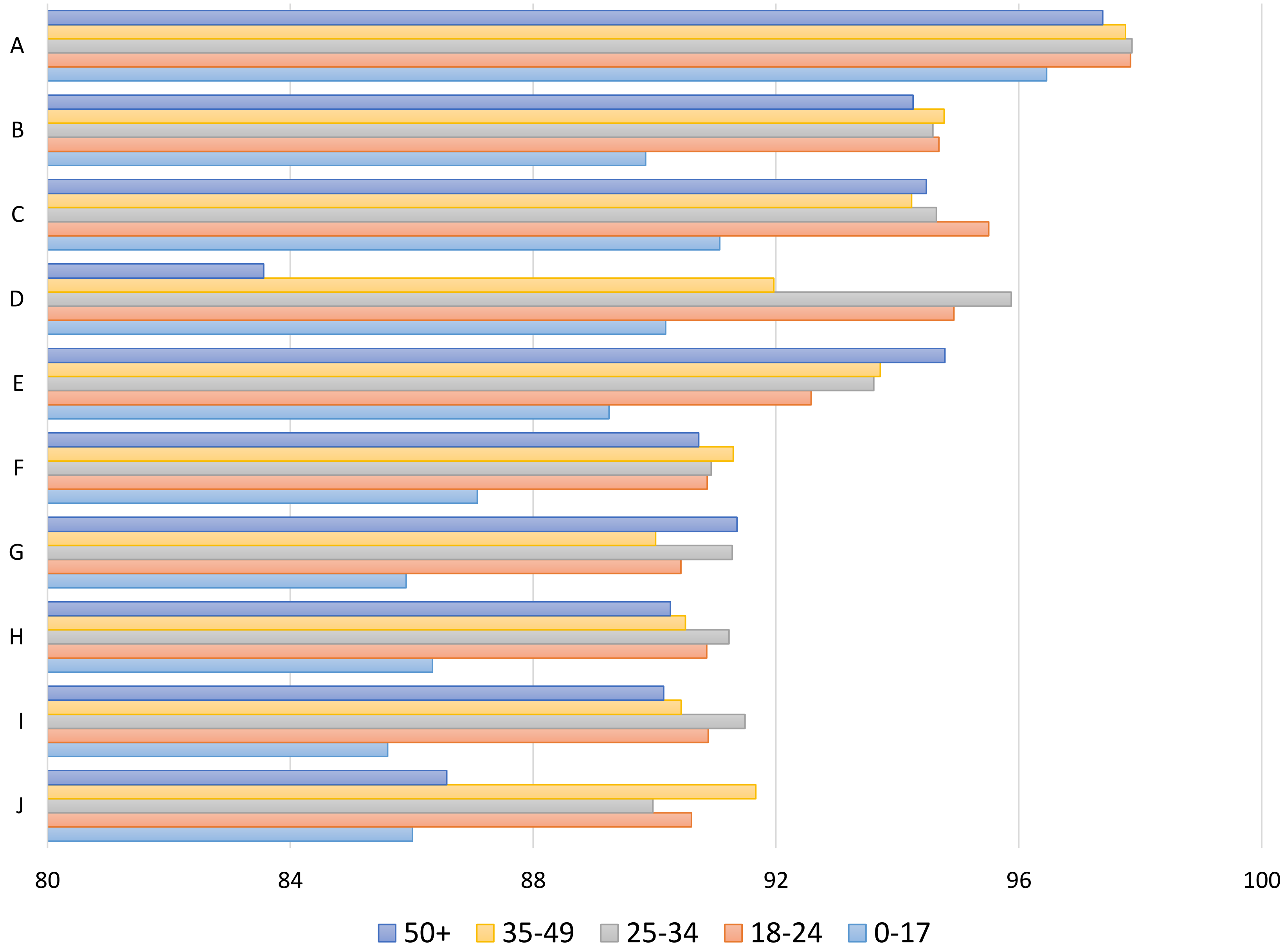
- The most effective AVs had the least variance between genders

- Ranking differs by gender



Anti-virus effectiveness by age group

- Most AVs struggled to protect 0-17 year olds
- Some vendors were inconsistent protecting 50+
- Ranking differs by age group

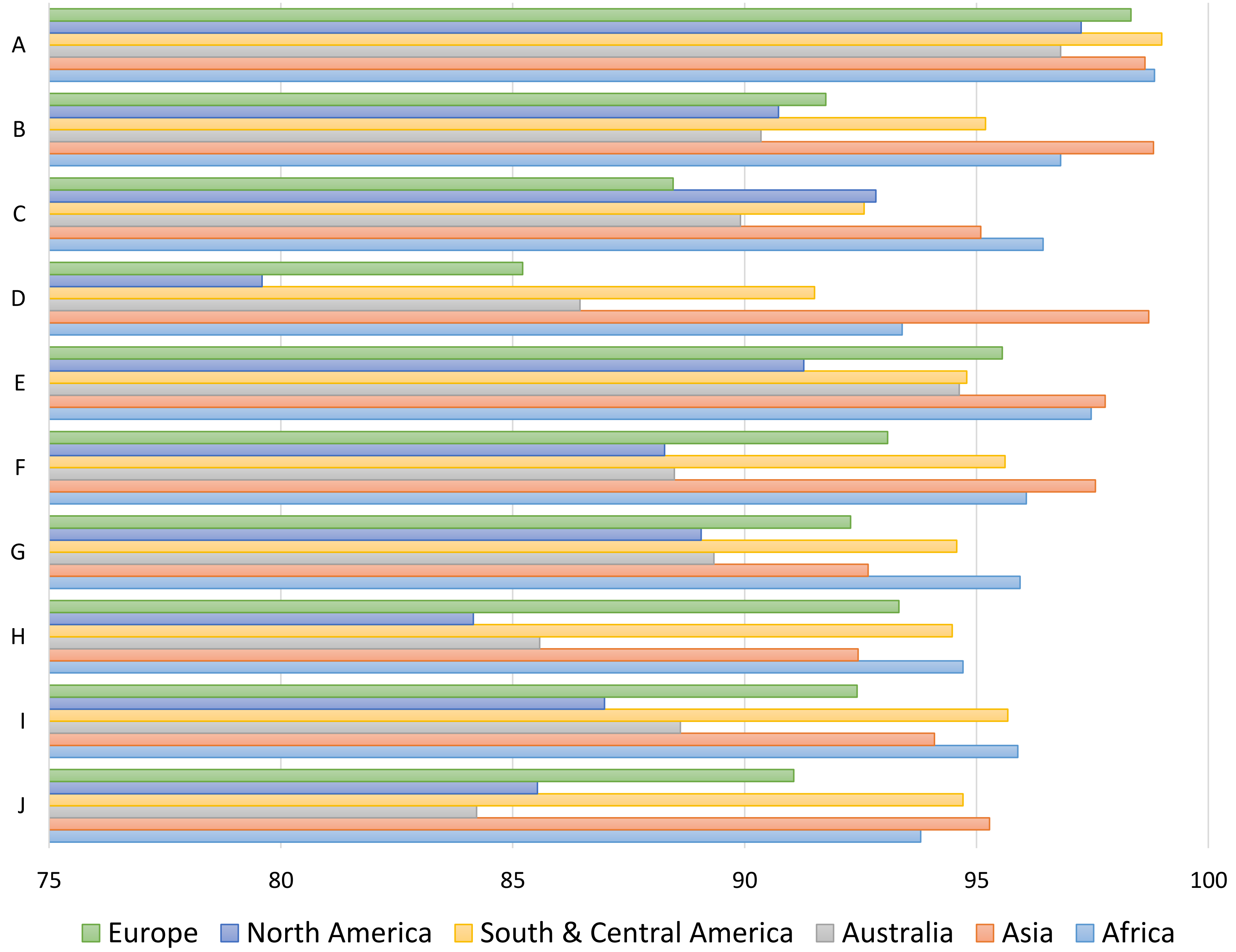


Anti-virus effectiveness by region

- North America had the lowest AVE for most vendors

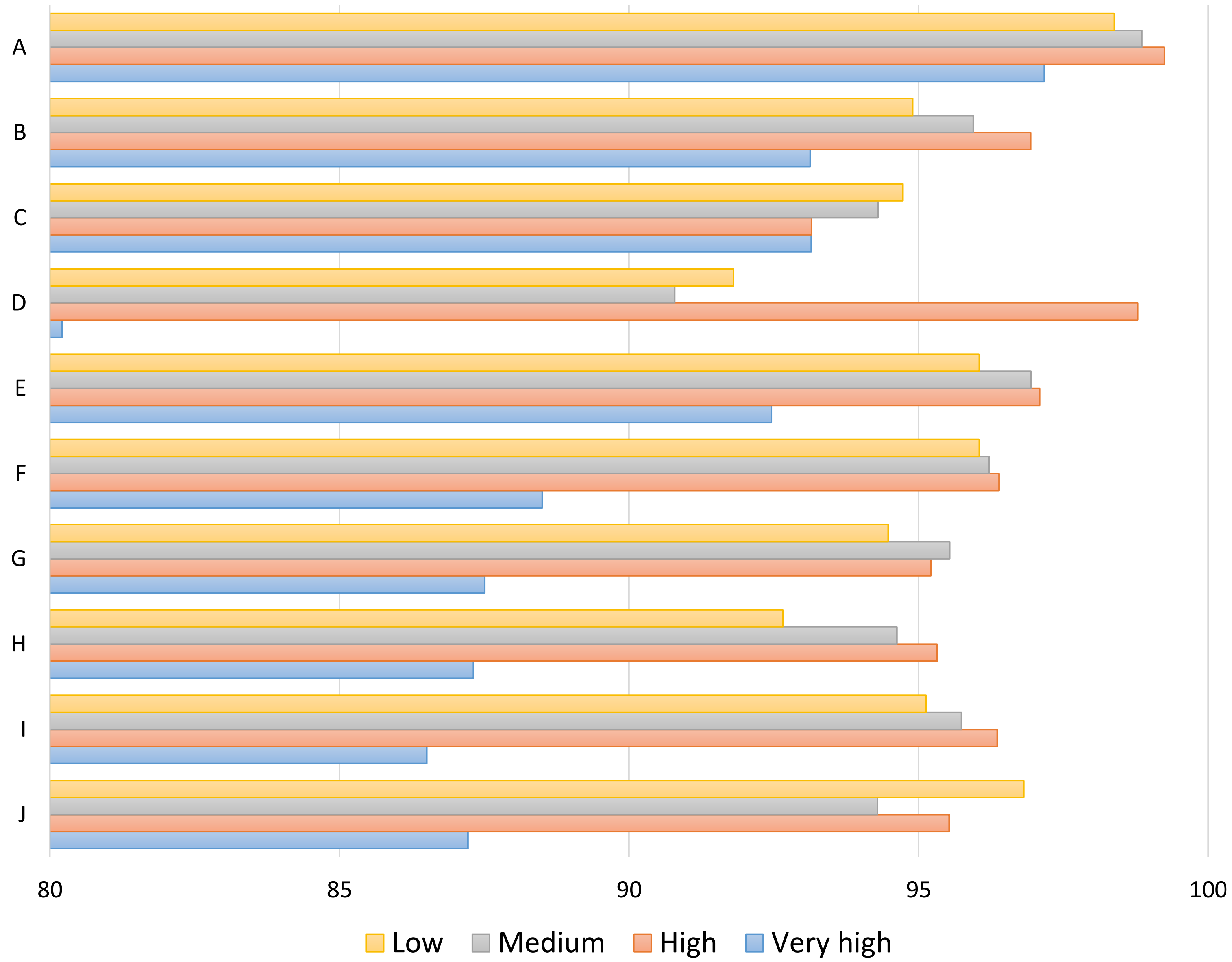
- North America had the highest vendor AVE variance

- Ranking differs by region



Anti-virus effectiveness by HDI category

- Very high had the lowest AVE for all vendors
- Very high had the highest variance in AVE (81%-97%)
- Ranking differs by HDI category



Key findings

AVE differs by malware types

- Classification differences between 3rd party AVs and MSRT
- Poor 3rd party AVs performance against unwanted software

AVE differs by user factors

- Differences in malware exposure and user behavior when faced with malware attacks
- Differences in attitude and behavior towards AV products

AVE differs by environmental factors

- Demographic differences
- Geographical differences in the malware landscape

Study limitations

- Only Windows 10 machines with known gender and age group
- Other factors may differentiate customers of 3rd party AVs
- MSRT families considered may not represent 3rd party AV priorities
- Comparison group AVE cannot be calculated

Future Work

For AV Research:

- Add data from other AV vendors to remove limitations
- Control for customer-based bias: clinical trials with randomly assigned AVs
- Conduct causality studies for differences in effectiveness
- Consider user behavior profiles (gamers, social networkers, etc.)
- Compare paid vs. free AVE

For AV vendors:

- Consider offering user-differentiated AV product

For AV testers:

- Complement lab tests with real-life measurements

Takeaways

- We live in a world of abundance of data; these kinds of tests are possible
- We can use real-life comparisons to measure effectiveness and drive improvement
- This was a hygiene test, and 3rd party AVE should be 100%. MSRT shouldn’t need to clean up infections when the AV is doing its job