Getting the most out of your testing

Quality Testing

Background

- Robert Neel
  - 25+ years
  - sysAdmin / netAdmin, LM on SBIRS, NSA as a GNEVA, USAA red team
- PEN Consultants
  - Offensive security
  - Small team (<10), but 3 are former nation-state hackers

Why This Topic Matters

- A penetration test is one of the most effective ways to validate the effectiveness of security controls and identify real-world vulnerabilities.
- BUT: “Penetration test” is subjective
- Most organizations are not getting “penetration test” level value
  - It’s just an automated scan, perhaps billed at a premium level
- “Penetration test” is used to describe things that barely qualify as a scan
  - A pentest in the length of time one MIGHT get thorough host/service discovery and enumeration.
- Industry is not helping -
  - HHS/HIPAA: “The [Department of Health and Human Services] estimates that each regulated entity would spend an average of 3 hours conducting penetration testing (with a low estimate of 2 hours and a high estimate of 10 hours)” - HHS pub. 2024-30983 (90 FR 898)
  - Android, iOS, & API pentest request: It’s Monday, we need it by the end of the week
  - DFW Airport: Hundreds of external IPs, 10s of thousands of internal IPs, black box (only external IP ranges provided), non-whitelisted (w/interference, treated as hostile), only “network tap” access provided, 5 days to complete.
- BOTTOM LINE: Odds are that you are not getting a real pentest, leaving you vulnerable.

Are You Even Ready for Testing?

- Baseline Security:
  - Actively managing and patching vulnerabilities
  - Basic security hygiene performed
  - Fixing common vulnerabilities first:
    - http://penconsultants.com/breachRiskAccept
    - https://penconsultants.com/PDFs/Presentations/TEEX_prez_2025.pdf
    - https://www.linkedin.com/posts/acook-cyber_pentest-activity-7249824057179713536-V-ct
  - For Red Teaming:
    - Technique Simulation: Must have logging and detections

        - Adversary Simulation: Must have your own SOC
- Implications of proceeding without that ^^^
    - Testing has a fixed amount of time/money:
        - All testing is limited in time
        - Limited by the client's authorized budget
    - The impact:
        - Test spends a lot of time verifying and reporting on the “basics”
            - No time left to look for more advanced things
        - Then again, no need to if script kiddie Metasploit can breach you
            - Bad guy not going to risk burning a zero day
- Maybe you need it to help influence the budget?
- Consider phases, with secure configuration upgrades in between:
    - Starting with a basic test - satisfy compliance, influence budget
    - Prepare for a year before moving to a standard level engagement
    - Prepare for another year before wasting money on a premium test

Definitions - What Are You *Actually* Buying?

- Vulnerability Scanning: Fully automated scanners looking for technical control gaps and misconfigurations in your systems - endpoints, network, web & mobile apps, cloud resources, etc.
    - Industry standard scanners: Nessus (network), Burp Suite Pro (web/api), etc.
    - NOT efficient to perform through a security stack
        - In fact, it’s close to worthless if blocking
    - NOT intended to test detection, mitigation, or response
        - BUT: It is LOUD, so many detections should trigger
    - 15-29% of firms call this a pentest - Source: Grok & ChatGPT
        - Even some top-10 firms call this penetration testing
        - “Next-gen”, “AI”, automated scanner “penetrates” vulnerabilities
    - We’ve had to label this as “Automated Testing” under penetration testing to compete.
- Vulnerability Assessment: Manual verification of what scanner(s) find
    - Manually “penetrate” suspected vulnerabilities found by the scanner
    - At most: Identifies vulnerabilities that an automated scanner is able to discover
    - Misses the most impactful vulnerabilities:
        - Most business logic vulnerabilities (e.g., A > C)
        - Exploit chaining (e.g., profile inject > PDF processor > SSRF)
        - Access control gaps (e.g., privilege escalation)
        - Sensitive data exposure
    - OFTEN referred to as a penetration test - Sources:
        - Our experience reviewing others’ SOWs and reports
        - OSINT says 60-80%!!!
            - Grok: “I estimate that 20-30% of all penetration testing firms (approximately 500-750 out of 2,500) do not place much emphasis on automated scanning but instead focus on manual testing that goes beyond anything a scanner can identify or exploit.“
            - ChatGPT: “Roughly 60–70% are mostly automated, meaning they rely heavily on scanners and only perform light manual verification. Conversely, 30–40% emphasize manual testing, with experienced pentesters going beyond what a scanner can identify or exploit.”
    - We had to label ours as “Basic Penetration Test” to compete

- Penetration Testing: Manual testing beyond what automation can discover
    - Note: This is still
        - ONLY looking for technical control gaps and misconfigurations in your systems
        - NOT efficient to test through a security stack in blocking mode
        - NOT intended to test your detection, mitigation, or response
    - We created three pentesting tiers:
        - Basic
            - Sub1 (mentioned above, mostly automated)
            - Sub2 (freestyle, very limited hours, mostly manual)
        - Standard (testing methodology comparable with reputable firms, 50/50 manual/automated)
        - Premium (no stone unturned, extensive manual)
    - Potential to give decent coverage of defensive solutions - firewalls, antivirus, EDRs, and so on
        - But not the primary focus
        - Depends on how the test is conducted, tier of service, etc.
- Red Teaming: Focused on testing detection, mitigation, and response of your security stack and people
    - Two primary branches:
        - Adversary Simulation - months long, stealthy, etc.
        - Technique Simulation - utilize common attacker tools and methods to verify detections
    - Other red teaming services:
        - IR Exercise: Often this is the end result of Adversary Simulation - “get noisy”. Sometimes it is more compressed (ex. a week of prep). But ultimately, this has the purpose to exercise IR - *train like you fight*. We also encourage very interactive purple teaming/side saddling red and blue during the process for optimal value and less burnout.
        - Tabletop Exercise: Perhaps we create a couple of injects in the network (breadcrumbs) that we explore during a tabletop walkthrough. Think of it more as us placing objects or activities in the network for illustration purposes during the tabletop, for demo purposes
        - Hunt: Gamification of blue team hunt training. We inject something into the network, announce the start of the game where blue looks for our activity, perhaps we give an increasing level of hints per day, until someone finds it and wins. Also has the benefit of finding other activities that are not us since the blue team has “eyes on” things.
- Pentesting vs Red Teaming: Mile wide, inch deep vs inch wide, mile deep
- NOTE: Definitions/service descriptions taken from:
    - https://penconsultants.com/pentestingTierComparison
    - https://penconsultants.com/serviceComparisonMatrix

Getting the Most Value

- Avoid Checkbox Testing: You’re wasting money; this is not security.
- Full white box, non-hostile testing
    - Everything you know, we know
    - Everything you have access to, we have access to
    - Whitelist the tester’s IP
    - Coordinate mitigations
    - *Discussed in-depth next section*
- Choose vendors who:
    - Offer white box testing
    - Work with you to prepare the environment
    - Allocate enough time

        - Full testing methodology and enumeration
        - Avoid gaps, DoS risks, locked accounts, overwhelming bandwidth, etc.
- Testing should be:
    - Planned outside of Q4
        - ~30-40% of tests are in Q4
        - Peak demand = sleep-deprived testers
        - Q2 is the best!
    - Long-running
        - 4 weeks at 20 hrs/week is better than 2 weeks at 40 hrs/week
        - Longer-running processes can run to completion
        - More time for creativity in attacks and recommendations
- Types of testing pair well together
    - Pentest + phishing campaign
    - Web app + mobile
    - Cloud + web | net
    - Wireless + Physical SE or assessment

White Box, Whitelisting, and Non-Hostile

- Goals:
    - Thoroughness - don't miss vulnerabilities
    - Time - relatively quick
    - Money - budget-friendly
- Give the testers everything
    - Full access and knowledge, user, dev, and admin credentials, logs, source code, etc.
    - "Giving this level of access is not representative of a real attacker."
        - *Which attacker do you want us to emulate?*
            - All of them?
            - Are decades of testing and millions of dollars authorized?
        - Then we need lots of access and knowledge to make up for that
    - Speeds up discovery, enumeration, and verification, and finds issues that would otherwise be missed (especially given the limited time)
- "Why do you need credentials?"
    - Low-level credentials are relatively easy to obtain, given enough time, and would fall under Social Engineering anyway
    - High-level credentials allow for thorough and quick testing
        - Search for role-based issues at ALL roles
        - Searching for buried resources that have weak permissions
        - Full AD/GP analysis
        - Full password audit vs slow, limited brute force
    - Metrics
        - Unauthenticated vs. user-level access = ~10x less thorough
        - Unauthenticated vs. admin-level access = ~20x less thorough
        - unauthenticated generally takes twice as long, costs twice as much, and is still not as thorough.
- Whitelisting reduces interference from security stack - firewalls, WAFs, EDRs, etc.
    - Provides better coverage and fewer false negatives (e.g., missing vulns)
    - Every minute that is spent bypassing defenses is a minute not spent looking for vulnerabilities
        - Months or years to run every test slow enough

        - Critical level finding if there is interference
        - If impactful enough, it is not practical or ethical to perform testing
    - Set to alert-only or monitor mode
        - Valuable to know if it would have detected/mitigated
    - After core testing
        - Non-whitelisted check performed on discovered vulnerabilities
        - Security stack tested separately or as part of red teaming
- Coordinate mitigations ahead of time
    - Only red teaming should be treated as hostile
    - False negatives
    - Critical level finding if being blocked
    - Waste time
- Multiple compliance standards require all of this as well
    - PCI DSS test is “fail” without whitelisting
    - A growing number of auditors demand:
        - full knowledge given
        - authenticated testing
        - no blocking
    - Quote from an auditor that constitutes a fail: “Non-credentialed testing, black box tests, scans, vulnerability assessments, vulnerability scanning that has been manually verified, crowdsourced or bug bounty pentests.”

Vendor Vetting and Red Flags

- What to look for:
    - Communication
        - What are your expectations?
        - Secure communication practices
        - Do they speak your language, in approximately your timezone, etc?
        - Informed progress during testing
    - Transparency
        - SOWs, pricing, methodology, and reporting
        - Open, honest, and not hiding behind intellectual property
        - Testing preparation
        - Timestamped notes and findings during testing
    - Clear deliverables & reporting:
        - Quality, prioritized risk ratings, explanations, screenshots, multiple mitigation recommendations, etc.
        - Reproducibility/PoC - preferably using only a browser or curl, no special tool requirements
        - Calls out third-party vulnerabilities as well!
            - “But I do not control that.”
            - “True, but you need to know so you can report it to the vendor, or find a new solutio.n”
    - Use of recognized testing methodology & frameworks:
        - PTES (for network pentest)
        - OWASP (for web or mobile app pentest)
        - MITRE ATT&CK (for red teaming)
        - NIST, OSSTMM, etc.

    - Cost
        - Significantly below average can signal
            - Deceptive marketing
            - Automated scan instead of real pentest
        - Above-average could be from
            - Inefficiencies
            - Markups
        - Network
            - Vulnerability scan: $2k+ external, $4k+ internal
            - Basic pentest: $4k+ external, $7k+ internal
            - Premium pentest: $10k+ external, $12k+ internal
        - Web
            - Basic: $7k+
            - Premium: $12k+
    - Education:
        - Help clients understand and improve
        - Explaining the pros/cons of various testing options
        - Process improvement recommendations
    - Protection
        - How is your data protected - during and after testing
        - Insurance - the standard professional & general liability, cyber, etc.
        - Contractually obligating themselves to everything
    - Vet their team
        - Like you're hiring them to some degree - because you are
        - Detailed bios, blog posts, code repos, social media, etc.
        - Combination of educated, trained, and certified, with diverse skills and experience
        - Assess their knowledge, compatibility with your team, honesty, trustworthiness, humility, etc.
    - References
        - Finding public references can be challenging - clients rarely approve
        - Ask for references - more clients are open to being connected with some directly
- What to avoid:
    - MSSP that manages the domain conducting the testing - conflict of interest, has a different mindset, etc.
    - Vague contracts & SOWs
    - Excessively low or high pricing without explanation
    - Scans sold as pentests
        - Actual quotes from SOWs and reports…
        - "Automated pentest"
        - "The scanner identified…"
        - "You should determine if this is a false positive."
        - "No attempts were made to exploit the findings."
        - "Vulnerabilities were identified using [tool name]"
    - "X hours of penetration testing before moving to assumed breach"
        - Time-based engagements are likely to leave gaps
        - Likely not completing a particular methodology
        - Thoroughness will be limited by the speed of the tester
    - No sample reports or:
        - Lacks details and quality
        - Includes tool output, especially with no analysis
        - Contains false positives (only a vuln scan has this)

        - Overstated severity

Switching & Alternating Vendors

- Changing vendors can offer new perspectives - but also carries risks
- Pros
    - Perspective: Might get a different skill set - but most firms will purposely put at least one new tester on the next engagement for this very reason.
    - Value: Potentially better value if the new firm is better.
- Cons
    - Perspective: New firm will not know history - spending more familiarization time and time on repeat/previous findings (less time finding new issues).
    - Value: New firm could be worse - there are more poor quality firms than good, so the odds are not in your favor.
- Recommendations:
    - Best: Alternate between only two trusted vendors every 6 months
    - Good: Stick with one, switch only when results stagnate
    - Risky: Rotate randomly - most vendors are low quality
- Key Metric: Quality vendors produce dozens of actionable findings
    - Assumption: you're paying for standard|premium level test, whitelisting, white box, not restricting your vendor, and providing all requested access/knowledge, etc.
    - Not just SSL/TLS weaknesses, header misconfigurations, user/acct enum vulns, etc.
        - Note: these are still important, though!
        - See https://penconsultants.com/PDFs/Presentations/TEEX_prez_2025.pdf
    - Drop the lowest competing vendors and replace
- Bad testing can be worse than no testing - it gives a false sense of security.

EXTRA: Example cost

Use as a baseline to know what you're getting

- Minimum Baseline (net testing)
    - Fully automated: ~$2k external, ~$4k internal
    - Basic pentest: ~$4k external, ~$7k internal
    - Premium pentest: ~$10k external, ~$12k internal
- 50+ live external endpoints (net testing)
    - Fully automated: ~$4k
    - Basic pentest: ~$7k
    - Premium pentest: ~$20k
- 2000+ live internal endpoints (net testing)
    - Fully automated: ~$6k
    - Basic pentest: ~$10k
    - Premium pentest: ~$30k
- Web app pentest (minimum)
    - Basic: $7k+
    - Premium: $12k+
    - Note: The number of features and roles drives up the cost
- Average-sized web app pentest:

    - Basic: ~$12k
    - Premium: ~$25k

EXTRA: Actual quotes from SOWs and reports

Red flags - Automated scans labeled as pentest:

- “Automated pentest”
- “The scanner identified…”
- “You should determine if this is a false positive.”
- “No attempts were made to exploit the findings.”
- “Vulnerabilities were identified using [tool name]”
- “Limited to unauthenticated scanning”
- “The tool was configured to…”
- “Completed a comprehensive scan”
- “Based on the scan results…”
- “Scanner detected [XSS | SQLi | etc.]”
- “We recommend further analysis to determine impact.”
- “Address only medium level risks and higher”

EXTRA: AI & LLMs

- AI is the new “next gen”
    - 10% useful
    - 90% hype
- Can AI replace manual pentesting?
    - Not anytime soon for sure
    - Three types of people/companies
        - They do not use AI at all: They are falling behind, and quickly
        - Uses AI and trusts it completely for core tasks: These people know just enough to be dangerous, run. AI can be horribly inaccurate, misleading, and even dangerous if not used responsibly
        - Using it with wrappers and safeguards placed around it: This is arguably the only intelligent approach
- “AI Penetration Testing” is a scam - just a new name for automated testing with some new LLM modules - run
- In general, the more “AI” is pushed in marketing and service descriptions, the more worried you should be.
- At the same time, if AI is not being used at all, the efficiency of the vendor will become increasingly worse (and therefore more expensive and less thorough)

EXTRA: Is all vulnerability scanning the same?

No!

We've spent over a decade refining our tools, methods, and configurations to extract real value from scans. Our scanning depth and coverage go well beyond what most vendors offer.

Take host discovery, for example. Nessus alone can not reliably discover hosts across all RFC 1918 IP space. Yet we do with our vulnerability scanning. That alone gives us visibility that others miss, often revealing shadow IT and forgotten assets that carry real risk.

Or port scanning: It is impractical for Nessus to scan all 65,535 TCP ports in any meaningful timeframe at scale. We do.

Even Nessus' own reporting on authenticated vs unauthenticated scanning coverage is deeply flawed - it reports a percentage based loosely on services, not hosts, and even that isn't accurate. We've developed methods for obtaining precise coverage metrics, helping us (and you) understand what's truly being assessed at a full, thorough level and what's not.