



ISO/IEC JTC 1/SC 27/WG 3 "Security evaluation, testing and specification"

Convenorship: UNE

Convenor: Bañón Miguel Mr



PWI_Evaluation_optimization_rev2

Document type	Related content	Document date	Expected action
Project / Other	Meeting: VIRTUAL 14 Sep 2026	2026-09-17	INFO

Terms of Reference for PWI

Optimizing Evaluation Methods Using Tools Based on New and Emerging Technologies

Introduction

ISO/IEC 15408 (multipart) establishes the model and criteria for evaluating security of IT products, including security functional requirements and security assurance requirements. ISO/IEC 18045 specifies the methodology for conducting such evaluations: evaluator actions, the evidence to be examined, and rules for reaching evaluation conclusions.

The Challenge

New tools and technologies can question the evaluation paradigm included in the standards. Several new evaluation methods or tools are described in the Annex.

Should the evaluation model specified in ISO/IEC 15408 (multipart) and ISO/IEC 18045 be changed? Should the focus be moved from the evidence provided by the Developer towards automated diverse conformity assessment methods and tools which will be completed by the final judgment made by the evaluator?

Expected outcome from the PWI

Standards developed in WG 3 have used proven assurance models over many years, but their underlying methods increasingly face challenges from modern software development practices, emerging technologies, and the growing use of AI in both product security and security evaluation. Many existing approaches are still rooted in relatively static, waterfall-oriented development and evaluation assumptions. The PWI therefore aims to identify methods that are no longer state of the art and to highlight important paradigms that are currently insufficiently addressed by existing standards. Attention should be given to areas where assurance increasingly depends on processes and complex supply chains rather than on the product itself.

The time-bounded PWI should investigate how AI, quantum computing, SBOMs/vulnerability analysis, and other new or emerging technologies can optimize evaluation methods. The key issue is to establish when a tool result can form part of the evaluation evidence; tools should support, rather than replace, the evaluator's responsibility for the conclusion.

Experts are asked to provide contributions on the following topics:

- technology taxonomy, a catalogue of use cases for the relevant standards, and an assessment of benefits, limitations, and risks. It could serve as a basis for selecting areas requiring further characteristics and as material for tool selection.

- guidance on qualification of technologies and on using related tool outputs as evaluation evidence. It should define minimum requirements for traceability of inputs, tool and model versions, configuration, repeatability, independent verifiability, control of false positives and false negatives, and protection of confidential evaluation data.
- proposals for standardized Evaluation Methods or Evaluation Activities for specific tasks under ISO/IEC 15408 and ISO/IEC 18045 with respect to evaluators', and developers' responsibilities..

This would enable laboratories to use new technologies consistently, in a controlled and verifiable manner, and to improve the efficiency and scalability of evaluations without reducing the required level of assurance.

Reference documents

[1] ISO/IEC JTC 1/SC 27/WG 3 N 3156 WG3_AI_Delegated_Evaluation_Contribution

Annex

Example Use Cases

For ISO/IEC 15408 and ISO/IEC 18045, the following technology groups could be particularly relevant:

- **Analytical, classification and agentic AI:** classification of test results and vulnerabilities, anomaly detection in logs, and prioritization of scenarios that need deeper analysis.
- **Agentic AI:** consistency and completeness analysis of documentation; traceability between requirements, implementation, and tests; and identification of discrepancies. An agentic system uses an LLM but extends it with access to tools, memory, rules, and a defined workflow. Under evaluator control, it can prepare a test, collect artifacts, and identify results for verification. Its factual correctness, repeatability, and risk of plausible but incorrect answers must be controlled.
- **Mature supporting techniques:** symbolic execution, SAT/SMT solvers, formal verification, fuzzing, and static and dynamic analysis. These provide measurable results for checking security properties, access control, information flows, protocols, and test coverage. AI can help plan, prioritize, and compare their results.

Other Potential Technologies

- **Digital twin:** a digital representation of the product under evaluation, its configuration and, where appropriate, cooperating components and network traffic. It supports test preparation, analysis of configuration changes, and reproduction of identified issues. GNS3, EVE-NG, and Mininet are examples of network-environment emulation tools that can be used to build part of such a representation.
- **Cyber range:** an isolated test environment that reproduces a topology, devices, services, and attack scenarios. It can use a digital twin, but its purpose is to conduct controlled exercises and tests, for example for updates, incorrect configurations, or attacks on network interfaces. It enables tests to be safely repeated, and their results compared across laboratories but does not replace testing of the actual TOE.
- **Knowledge graph:** a structured model of relationships, for example: "requirement -- component -- test -- result -- vulnerability -- evidence". It enables an evaluator to quickly identify which evidence supports a requirement and which tests or components are affected by a new vulnerability. Such relationships can be stored in a graph database, for example Neo4j.
- **Supply-chain analysis automation:** an SBOM is a structured inventory of software components, while VEX communicates whether a known vulnerability affects a specific product. Correlating these with vulnerability databases helps identify components requiring analysis and track the effects of updates. Example formats include SPDX and CycloneDX for SBOMs, and OpenVEX and CSAF for VEX;

- **Quantum computing as a basis for evaluation methods** can be considered if quantum or quantum-inspired algorithms improve the selection of test scenarios, fuzzing campaign planning, or dependency-graph analysis. Their results should not be accepted as stand-alone evaluation evidence unless their quality, stability, and repeatability have been demonstrated.

All named products are examples of tool classes only. The PWI should define requirements for every tool used, particularly for repeatability, configuration traceability, and independent verification of results.