Project Title	Civil Cyber Range Platform for a novel approach to cybersecurity threats simulation and professional training
Project Acronym	CYBERWISER.EU
Project Number	786668
Type of instrument	Innovation Action
Topic	DS-07-2017 Cybersecurity PPP: Addressing Advanced Cyber Security Threats and Threat Actors
Starting date of Project	01/09/2018
Duration of the project	30
Website	www.cyberwiser.eu

D4.2 Real-time Performance and Evaluation Criteria

Work Package	WP4 Training material, scenarios and evaluation
Lead author	Aida Omerovic (SINTEF)
Contributors	Gencer Erdogan (SINTEF), Åsmund Hugo (SINTEF), Dario Varano (UNIPI), Gianluca Dini (UNIPI), Cinzia Bernardeschi (UNIPI), Gigliola Vaglini (UNIPI), Liliana Ribeiro (EDP), José Lourenço (EDP), Giorgio Aprile (FFSS), Ioannis Kechaoglou (RHEA), Antonio Álvarez (ATOS), Anže Žitnik (XLAB), Manca Bizjak (XLAB)
Peer reviewers	Valerio Vitangeli (FFSS), Raniero Rapone (AON)
Version	V1.0
Due Date	31/08/2019
Submission Date	30/08/2019

Dissemination Level:

X	PU: Public
	CO: Confidential, only for members of the consortium (including the Commission)
	EU-RES. Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)
	EU-CON. Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)
	EU-SEC. Classified Information: SECRET UE (Commission Decision 2005/444/EC)

The work described in this document has been conducted within the CYBERWISER project. This project has received funding by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 786668.

Version History

Revision	Date	Editor	Comments
0.1	20/02/2019	Aida Omerovic (SINTEF)	Initial version of the deliverable. Initial outline.
0.2	01/04/2019	Dario Varano (UNIPI)	Contribution to section 3.1.2
0.3	04/04/2019	Dario Varano (UNIPI)	Contribution to section 3.3 on the Bloom's Taxonomy
0.4	12/04/2019	Dario Varano (UNIPI)	Adding content to section 3.3.1
0.5	23/04/2019	Liliana Ribeiro (EDP)	Adding content to section 3.1.3
0.5	23/04/2019	Ioannis Kechaoglou (RHEA)	Adding content to section 3.1.1
0.6	10/05/2019	Aida Omerovic (SINTEF)	Adding contents to 3.3
0.7	21/05/2019	Antonio Álvarez (ATOS)	Revision of content produced up to now
0.8	10/06/2019	Anže Žitnik, Manca Bizjak (XLAB)	Adding contents to 3.2.
0.9	17/06/2019	Anže Žitnik, Manca Bizjak (XLAB)	Adding contents to 3.2.
0.10	21/06/2019	Aida Omerovic (SINTEF)	Proposed and added templates for indicator and evaluation criterion specification, to Section 4.2
0.11	08/07/2019	Gencer Erdogan (SINTEF)	Wrote Section 2. Checked all comments provided so far.
0.12	22/07/2019	Åsmund Hugo (SINTEF)	Wrote Section 3.3 apart from Blooms Taxonomy. Including RHEA branched version of 0.11
0.13	30/07/2019	Åsmund Hugo (SINTEF)	Section 2.2 draft completed with indicators and criteria from UNIPI and EDP.
0.14	01/08/2019	Giorgio Aprile (FFSS)	General review of the document, contributions added to chapters 3 and 4.
0.15	05/08/2019	Åsmund Hugo (SINTEF)	General formatting
0.16	05/08/2019	Gencer Erdogan (SINTEF)	Methodology of identifying and specifying indicators and performance criteria, added to chapter 4.
0.17	06/08/2019	José Lourenço (EDP)	Extension of indicators and criteria in section 5.2
0.18	08/08/2019	Antonio Álvarez (ATOS)	Input to sections 3.2.1, 3.2.2, 3.2.4. Addition of some comments along the document
0.19	09/08/2019	Anže Žitnik (XLAB)	Input to sections 3.2.1, 3.2.2, 3.2.4. Addition of some comments along the document
0.20	06/08/2019	Åsmund Hugo (SINTEF)	General formatting and extension of assumptions in chapter 4. Input from Antonio taken into consideration.
0.21	13/08/2019	Gencer Erdogan (SINTEF)	Reviewed the whole document. Revised Sections 1 and 5. Added overview figures for all evaluation criteria and indicators in Section 5. Wrote Section 6. Checked and corrected the formatting of text, tables, figures, etc. in the document.
0.22	23/08/2019	Gencer Erdogan (SINTEF), Raniero Rapone (AON)	Updated the document according to the first internal review performed by AON.
0.23	27/08/2019	Gencer Erdogan (SINTEF), Valerio Vitangeli (FFSS)	Updated the document according to the second internal review performed by FFSS. Final refinements.

1.0	30/08/2019	Gencer Erdogan (SINTEF), María Teresa García González (ATOS)	Updated the document according to the Quality Assurance performed by ATOS. Document ready for submission.

List of Contributors

The list of contributors to this deliverable are presented in the following table:

Section	Author(s)
Executive Summary	Aida Omerovic (SINTEF), Gencer Erdogan (SINTEF)
1	Aida Omerovic (SINTEF), Gencer Erdogan (SINTEF)
2	Gencer Erdogan (SINTEF)
3	Åsmund Hugo (SINTEF), José Lourenço (EDP), Ioannis Kechaoglou (RHEA), Antonio Alvarez (ATOS), Dario Varano (UNIPI), Gianluca Dini (UNIPI), Cinzia Bernardeschi (UNIPI), Gigliola Vaglini (UNIPI), Pericle Perazzo (UNIPI), Anže Žitnik (XLAB), Manca Bizjak (XLAB)
4	Aida Omerovic (SINTEF), Gencer Erdogan (SINTEF), Dario Varano (UNIPI), Ioannis Kechaoglou (RHEA), Antonio Alvarez (ATOS), Anže Žitnik (XLAB)
5	Dario Varano (UNIPI), Giorgio Aprile (FFSS), José Lourenço (EDP), Gencer Erdogan (SINTEF)
6	Gencer Erdogan (SINTEF), Åsmund Hugo (SINTEF)

Keywords

Evaluation criteria, performance evaluation, real-time performance evaluation, cyber-range participant performance evaluation, evaluation indicators, method, guideline, trainee performance, state of the art, state of the practice.

Disclaimer

Table of Contents

List of figures

List of tables

Executive Summary

Task T4.1 of CYBERWISER.eu provides training material in the form of courses at various levels of the CYBERWISER.eu Platform offer. Some of the courses are decomposed into modules. Each course is associated with meta-information about its goals, objectives, difficulty level, duration, etc. A detailed account of the courses and their contents is provided in Deliverable D4.1 [12]. Currently, there have been developed a set of courses for the two offering levels Primer and Basic. Courses for higher advancement levels (Intermediate and Advanced) are at the time of writing under development and will be documented in Deliverable D4.4 (due M18, February 2020). The courses will be supported by training scenarios developed in Task T4.2. The evaluation criteria developed in Task T4.3 will be used to evaluate the progress of the trainees in the scenarios.

Task T4.3 of CYBERWISER.eu provides three main artefacts:

- a method for identification and specification of the evaluation criteria to evaluate the performance of trainees (course participants),
- a set of specific evaluation criteria to be applied in the CYBERWISER.eu courses which will also be used in the pilots defined in WP5,
- a state-of-practice and state-of-the-art with respect to performance evaluation.

The abovementioned artefacts are described in this report. The set of evaluation criteria have been developed considering the learning goals and objectives of the exercises for the pilots in WP5 (defined in Task T5.1 and reported in Deliverable D5.1) and the expected corresponding courses to be developed in Task T4.1 for the Intermediate and Advanced offering levels. There are two main reasons to this:

1. The pilots in WP5 are interested in certain learning goals and objectives that require the availability of the technical assets: Performance Evaluator, Economic Risk Evaluator, Countermeasure Simulator, and Vulnerability Assessment Tools.
2. According to the CYBERWISER.eu learning path defined in Deliverable D4.1 [12], courses in which the above technical assets are used will be available in the Intermediate and Advanced offering levels.

Evaluation criteria for the courses beyond those that will be used by the pilots in WP5 and that require the abovementioned technical assets will be developed as part of the validation activities of CYBERWISER.eu (using the method for identifying and specifying evaluation criteria documented in Section 4).

The trainer and the platform shall use the relevant evaluation criteria to extract and aggregate the information needed for evaluation of the progress of the trainee.

With respect to evaluation criteria, we have defined:

- in total 8 evaluation criteria and 12 indicators for Pilot 1, academic training,
- in total 2 evaluation criteria and 4 indicators for Pilot 2, transport infrastructure,
- in total 2 evaluation criteria and 5 indicators for Pilot 3, energy infrastructure.

1. Introduction

This section provides an introduction in terms of purpose, relation to other work in the project and the scope of this deliverable, structure of the document, and a glossary of acronyms.

1.1 Purpose

Task T4.3 (real-time performance and evaluation criteria) provides a baseline for evaluation of the performance of trainees during cyber-range exercises/scenarios as well as to assess the current knowledge and skills of trainees. Throughout this document we use the terms exercises and scenarios interchangeably when the distinction is not important.

There are three main artefacts developed in Task T4.3 described in this document. Firstly, the task designs a method for identification and specification of the so-called "evaluation criteria", which are linked to the course-specific "learning objectives". The method enables, through a structured process guideline and a set of specification templates, maintenance of the proposed criteria and development of the future ones (for existing or new courses).

Secondly, the task defines a set of specific evaluation criteria to be used in the courses to be developed for the Intermediate and Advanced offering levels, which are also relevant for the pilots in WP5. There are two main reasons to this:

- The pilots in WP5 are interested in certain learning goals and objectives that require the availability of the technical assets: Performance Evaluator, Economic Risk Evaluator, Countermeasure Simulator, and Vulnerability Assessment Tools.
- According to the CYBERWISER.eu learning path defined in Deliverable D4.1 [12], courses in which the above technical assets are used will be available in the Intermediate and Advanced offering levels.

The identified set of evaluation criteria described in this document are therefore used to capture the response of trainees during cyber-range exercises with a particular focus on the learning objectives of the courses relevant for the pilots. The identified evaluation criteria will also be used (and adjusted if necessary) in exercises assessing the degree to which risk models selected and/or developed by the trainee reflect the cyber-risk posture of the target system.

The method for identifying and specifying evaluation criteria supports specification of new evaluation criteria within any one of the planned courses within the context of CYBERWISER.eu. The method, the templates and the specific evaluation criteria are founded on the relevant state-of-the-art and in particular on the expertise and experience of the partners within the CYBERWISER.eu consortium.

The state of the practice and state-of-the-art reported in Section 3 represent collectively the third artefact, which is an additional contribution in addition to the abovementioned method and set of evaluation criteria.

Finally, in this document, we refer to other deliverables whenever necessary and do not repeat the content from the referred deliverables in this document. For example, the identified evaluation criteria are documented in Section 5, while the related exercises on which the evaluation criteria will be applied are documented in detail in Deliverable D5.1.

1.2 Relation to other work in the project and the scope of this deliverable

CYBERWISER.eu has three business priorities:

- To **define capacity building paths** for cybersecurity professionals in Europe;
- To **decrease barriers** to entry to sophisticated and mainstream cybersecurity competences for ICT-intensive organisations across disciplines and business sectors;
- To contribute to **raising awareness** around cyber-risks and cyber-range.

To best address the above-mentioned business priorities, the entire business proposition of CYBERWISER.eu is sliced and diced around four progressive levels of training and information which is reflected in the CYBERWISER.eu Platform offer. These levels are depicted in Figure 1.

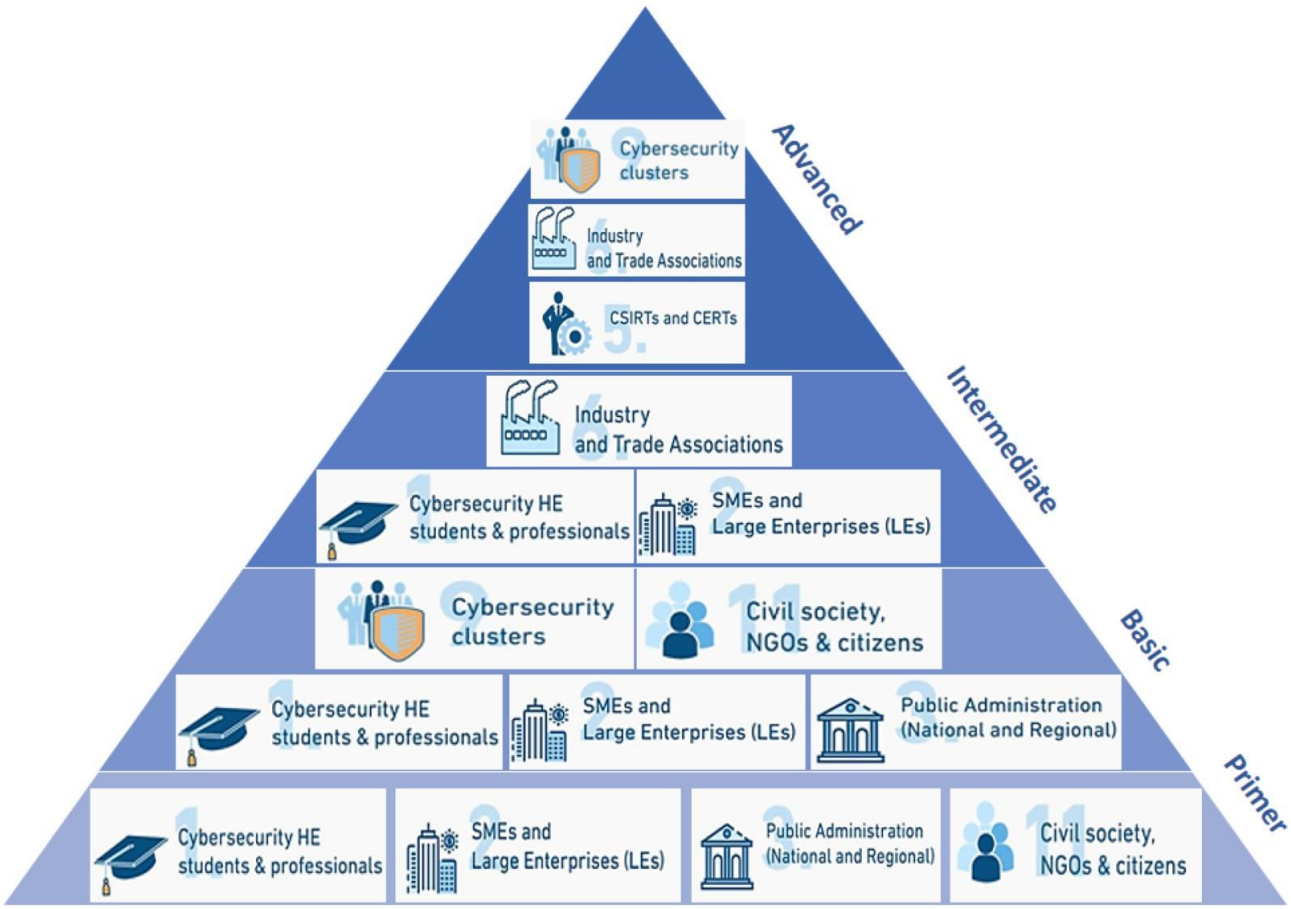

Figure 1. CYBERWISER.eu “pyramidal offer”

WP4 of CYBERWISER.eu provides training material in the form of courses at all the four levels. The courses are decomposed into modules. Each course is associated with meta-information about its goals, objectives, difficulty level, duration, etc. A detailed account of the courses and their contents is provided in Deliverable D4.1 [12]. Currently, there have been developed a set of courses for the two offering levels Primer and Basic. Courses for higher advancement levels (Intermediate and Advanced) are at the time of writing under development.

This deliverable reports on the results of Task T4.3 of CYBERWISER.eu. As indicated in Section 1.1, Task T4.3 has developed performance evaluation criteria and supporting indicators in order to facilitate the evaluation of progress of the trainee with respect to educational objectives of the exercises defined in WP5 and the corresponding expected courses in Task T4.1 and cyber-training scenarios in Task T4.2 (i.e. scenarios, as in exercise is equivalent to a scenario) in question. This includes the evaluation of trainees in terms of selected/developed risk models, as well as risk mitigation options provided by the countermeasure simulator (described in Deliverable D2.4 as well as Section 3.2.4). The trainer and the platform shall use the relevant evaluation criteria to extract and aggregate the information needed for evaluation of the progress of the trainee. The method, the templates (in Section 4) and the specific evaluation criteria (in Section 5) are founded on the relevant state of practice and state-of-the-art.

Hence, this deliverable provides three main artefacts: a method for identification and specification of the evaluation criteria, a set of specific evaluation criteria for the CYBERWISER.eu courses, and a state-of-the-art including state of the practice as reported by the partners in CYBERWISER.eu.

The clear and accurate definition of the evaluation criteria is paramount for the development of the internal modules of the Performance Evaluator (PE). The PE is a key component of the CYBERWISER.eu architecture, in charge of assessing how well the participants in the training are performing. Deliverable D2.5, produced in the design Task T2.2, documents the internal architecture of the PE. In particular, there is a module called Evaluator, in which the internal algorithm needs to be based on the evaluation criteria developed in Task T4.3 and later stages of the project. These criteria are the drivers for the internal implementation activities to be carried out in this module in the context of Task T2.4.

1.3 Structure of the document

The document is structured as follows:

- Section 1 (this section) provides an introduction in terms of purpose, relation to other work in the project and the scope of this deliverable, structure of the document, and a glossary of acronyms.
- Section 2 explains the process leading to the artefacts reported in this document.
- Section 3 provides the state of the practice (as reported by the partners in CYBERWISER.eu) and state of the art.
- Section 4 describes the guidelines for identifying and specifying criteria for the evaluation of trainee performance.
- Section 5 documents the established evaluation criteria and indicators.
- Section 6 provides conclusions and final remarks.

1.4 Glossary of Acronyms

Acronym	Description
ADR	Anomaly Detection Reasoner
ATC	Academic Training Criterion
ATI	Academic Training Indicator
BAN Logic	Burrows–Abadi–Needham logic
BYOD	Bring Your Own Device
CNAIP	National IT crime protection centre for critical infrastructure
CS	Countermeasure Simulator
CSFR	Cross-Site Request Forgery
CTF	Capture the Flag
DDoS	Distributed Denial of Service
DNS	Domain Name System
DoS	Denial of Service
ERE	Economic Risk Evaluator
ETC	Energy Training Criterion
ETI	Energy Training Indicator
GQM	Goal-Question-Metric
ICT	Information and Communications Technology
IP	Internet Protocol
IT	Information Technology
LAN	Local Area Network
LEAP	Liberal Education and America's Promise
OS	Operating System
OSSEC	Open Source HIDS SECurity
PCAP	Packet Capture
PE	Performance Evaluator

Acronym	Description
PII	Personally Identifiable Information
SQL	Structured Query Language
TTC	Transport Training Criterion
TTI	Transport Training Indicator
USB	Universal Serial Bus
VALUE	Valid Assessment of Learning in Undergraduate Education
WISER	Wide-Impact cyber SEcurity Risk framework
WP	Work Package
XSS	Cross-site scripting
XXE	XML External Entity (XXE) injection

Table 1. Table of acronyms

2. The process leading to the artefacts

Figure 2 illustrates the process carried out to produce the artefacts described in this document. In Step 1, we defined a method to systematically identify the evaluation criteria. In Step 2, we described the state of practice as currently carried out by the partners in CYBERWISER.eu, as well as the state of the art on evaluation of participants in courses, mainly focused on related methods and theories in the literature. Finally, in Step 3, we identified evaluation criteria using the method defined in Step 1. The evaluation criteria defined in Step 3 support the evaluation of participants in the exercises (developed in Task T4.2) that are part of the courses developed in Task T4.1. However, to facilitate the pilots in WP5, the evaluation criteria also take into consideration the pilot needs in terms of evaluation of participants in the pilot-specific exercises.

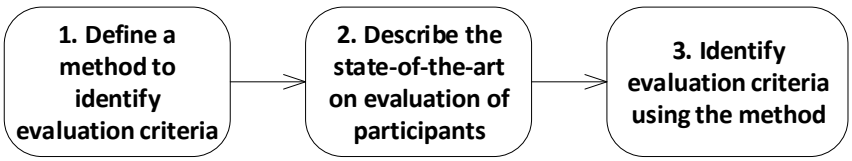

Figure 2. The process leading to the artefacts

Although the steps of the process in Figure 2 are illustrated sequentially, the Steps 1 and 2 were partly carried out in parallel. While obtaining the state of the art in Step 2, we updated the method in Step 1 if found necessary, to be in line with the current practice. The method for identifying evaluation criteria consists of two overall steps. In the first step, we describe evaluation criteria using a predefined template. In the template, we describe amongst others an exercise in which the evaluation criterion will be used, and we also indicate the course (developed in Task T4.1) in which the criterion is relevant. In the second step, we define for each evaluation criterion, one or more evaluation indicators. An evaluation indicator describes one concrete measurement approach to help evaluate the extent to which the evaluation criterion is fulfilled. Section 4 provides a detailed description of the method and the templates used for describing evaluation criteria and their supporting indicators.

As pointed out above, the state of the art carried out in Step 2 consists of two parts. The first part is related to current practice in terms of performance evaluation as reported by the partners in CYBERWISER.eu. This includes the current practice described by the partners providing pilots (UNIPI, EDP, and FFSS), as well as the current evaluation practices carried out by RHEA. With respect to current practice, the following approaches are described: Capture the Flag (CTF), sensor-based (automated), questionnaires, evaluation through academic projects and tests, and surveys. Additional measurements are also taken via attack simulators, vulnerability assessment tools, as well as custom detectors of specific actions, amongst others. With respect to state-of-the-art methods and theories (the second part), we look closer into Bloom's Taxonomy, Goal-Question-Metric, and Privacy Scorecard. Section 3 describes the state of the practice and state of the art.

The evaluation criteria identified as part of Step 3 are documented in Section 5. Each evaluation criterion is described using the evaluation criterion template, and for each evaluation criterion we also describe one or more indicators. The indicators are described using the indicator template.

3. State of the practice and state of the art

This section describes the state of the practice in terms of performance evaluation as reported by the partners in CYBERWISER.eu (Sections 3.1 and 3.2), and the related state of the art methods and theories (Section 3.3).

3.1 Current practice in terms of performance evaluation, as reported by the CYBERWISER.eu consortium partners

This section describes the current practice in terms of performance evaluation by the relevant partners (RHEA, UNIPI, EDP, FFSS) of CYBERWISER.eu:

- Current practices by RHEA cover Capture the Flag (CTF), automated sensor-based evaluations, and questionnaires.
- Current practices by UNIPI cover academic projects, written tests, practical tests, oral tests.
- Current practices by EDP cover surveys and interactive practical exercises.
- Current practices by FFSS cover video-based training with a learning test at the end.

3.1.1 Current evaluation practices by RHEA

To capture and evaluate the performance of the trainee RHEA has identified the following three main evaluation methods:

- Capture the flag (CTF)
- Sensor-based (automated)
- Questionnaire

Following a description of each method and a comparison table with their pros and cons at the end of the section.

Capture the flag (CTF)

This method requires from the trainee to find some artefacts during the training and submit them in an online platform for evaluation.

The artefacts (called flags) are planted during the preparation of the training in proper places or are identified and extracted from the available data. Each of the flags must signify an event or an action important to the evaluation of the trainee. Examples of flags are:

- A text in a file inside the root directory of a Linux machine.
  The text can be read only from a user with root privileges thus it signifies that the trainee managed to access the machine and escalate its privileges to root.
- An encryption key inside a ransomware.
  Finding the key inside the malware will indicate that the trainee has the competency to reverse engineer it and identify/extract the key.
- A protocol parameter value in a specific packet (PCAP).
  Identifying the value indicates knowledge of network traffic analysis.

This method is supported by an online platform where the trainee can submit their flag. The platform will check them and automatically grade the trainee. Some common extra features of these kinds of platforms are the ability to provide hints to the trainees (with or without a point penalty) and the division of the contest in levels requiring a percentage of completion of one level before going to the next.

The implementation is reasonably easy in that it requires effort by the trainer only during the preparation of the training and has a minimal effect to the simulated environment. Moreover, the trainee is experiencing only a minimal interruption at the submission of the flag.

The flags cannot cover all the quantitative metrics (e.g. website defacement) and they cannot cover qualitative metrics at all. Their implementation is best suited for Red team trainings (such as ethical hackers) [11][13] where the trainee penetrates a network/application and finds the flags in its way.

Sensor-based (automated)

This method requires the installation and configuration of sensors capable to detect evaluation indicators (e.g. user-profile creation and web site down time). The data then are sent and stored in a central database for processing. The result of the processing is the performance evaluation (score) of the trainee.

In this context, there are two types of sensors: agents and virtual machines. The agents are installed in the endpoints, as kernel modules or user-space applications monitoring the status of the endpoint end its connectivity. The virtual machines are deployed across the network gathering information about the status of the network and the availability/quality of the network services available.

The implementation is complex and requires not only a lot of extra effort during the preparation of the training but also monitoring during the execution. The sensors exist in the same environment as the trainees causing interference that must be clearly communicated to them beforehand. The proper operation of the sensors as well as their protection from the trainees is crucial for the evaluation.

With this method the trainee is not interrupted at all during the training and the evaluation is fully automated. Another significant aspect is that the evaluation can be based only on the end results allowing the trainee to create their path to success promoting out-of-the-box thinking. On the limitation of the method is the inability to measure qualitative metrics.

Both Blue and Red teams can be trained with this method.

Questionnaire

The questionnaire method is similar to the CTF method with respect to the way it gathers information from the trainee in a question-answer form.

The questionnaire can be completed during or at the end of the training gathering quantitative and qualitative data (logic-related metric and situational awareness) making this the only method to cover both. The qualitative metrics introduce the need for a human to process the responses, most of the times. It is possible, with a well-prepared questionnaire, to capture qualitative data allowing also automated evaluation but this requires significant effort during the preparation, and it is most likely that some cases will not be covered.

Answering the questions during the training implies interruption of the training and in some cases (mostly for qualitative metrics) significant. On the other hand, answering questions after the training may result in loss of information. The questionnaire method fits well for blue and red team trainings and it is very simple to implement with the benefit of not interfering with the simulated environment.

Method comparison

Table 2 provides a comparison of the abovementioned approaches in terms of pros and cons.

	Pros	Cons
Capture-The-Flag	• Minimal training interruption • Reasonably easy implementation (plant flags + Boolean server-side evaluation) • Does not interfere with the simulated environment	• Requires trainees to "submit" their findings • Does not cover all quantitative metrics (like web site defacement) • Does not cover qualitative metrics • Works best for red-team trainee's evaluation
Sensor-based (automated)	• Covers result-driven metrics • Does not require human interaction • Works for both BLUE and RED team trainees' evaluations • Does not interrupt the training	• Does not cover all qualitative metrics (situational awareness, decisions) • Complex implementation • Requires monitoring infrastructure • Interferes with the simulated environment

Questionnaire	• Covers also qualitative metrics like
	  ○ Logic-related metrics (e.g. decision reasoning)
	  ○ Situational awareness
	• Covers also qualitative metrics
	• Works for both BLUE and RED team trainees' evaluations
	• Does not interfere with the simulated environment
		• Significant training interruption
		• Non "real-time" performance evaluation
		• Requires human interaction (trainer and trainee)

Table 2. Method comparison (CTF, sensor based, and questionnaires)

3.1.2 Current evaluation practices by University of Pisa

The current practice for performance evaluation in the University of Pisa (UNIPI) follows the Italian criteria in terms of methods for testing trainees' learning capabilities. The standard method for trainees' evaluation consists of:

- Academic Project;
- Written test;
- Practical Test;
- Oral test.

Not all evaluation techniques are mandatory; depending on the course only a subgroup can be chosen. A description of each kind of test will follow.

The aim of the Academic Project is usually to implement an ICT system or a sub-system, which fits the course learning objective. The evaluation is course-dependant, in general it is a check that the implementation is compliant with the initial specifications. The result can be Boolean for the access to the other tests or giving additional points to the fine score.

The written test is usually composed by a list of exercises to be solved. The aim is to test the ability of the trainees on analysing, describing or dimensioning a system's component. It can be made of open answer or multiple choices.

The practical test can be made of a list of exercises or a unique exercise composed by different sub-exercises to be solved on a PC. The type of the exercises is course-dependant, an example can be the implementation of a software functionality using a programming language. The aim is to check the knowledge of the trainees on the practical content of the course.

The oral test is almost always present in all courses. It is composed by a set of questions about the course's content. Usually, each question covers a macro-section of the course. The aim is to evaluate the knowledge of the trainees on theoretical concepts, along with the public speaking ability.

To improve the explanation of how an academic project is evaluated, an example will follow. The example considers a typical academic project for the cybersecurity course of the Department of Information Engineering of the University of Pisa. The assignment for the project is as follows:

- Topic: Developing of a small security application;
- Brief description: Develop an application employing a key-establishment protocol for agreeing on a symmetric session key and a session secured with such a key;
- Programming language: C, Java, Java for Android, Python or C#;
- Requirement: Test the protocol with the BAN Logic, ensuring its robustness in the studied scenario.

The evaluation of such project consists of verifying the following requirements:

- Correctness of the general security design:
  - no flaws should be present in the application design.
- Correctness of the key-establishment protocol:
  - no flaws should be present in the protocol;
  - the BAN logic is required to be present;
  - the candidate should highlight and discuss the a-priori hypotheses of the protocol;
  - the candidate should prove these a-priori hypotheses to be reasonable for the specific application; offering the property of perfect forward secrecy is also a bonus.
- Correctness of the session protocol:
  - No flaws should be present in the session protocol;
  - the protocol should resist against classic attacks like eavesdropping, man-in-the-middle, message injection, message replay;
  - the resistance against advanced attacks like padding oracle attack is also a bonus.

3.1.3 Current evaluation practices by Energias de Portugal

EDP – Energias de Portugal, S.A. is a Portuguese organization with presence in several geographies. EDP has about 12.000 employees, and the need to make their employees aware of common cybersecurity risks they may be exposed to is one of the main priorities of EDP.

Annually, EDP conducts a survey to evaluate the cybersecurity awareness of its employees. With these, it is possible to ascertain the efficiency of internal training programmes. One the most significant trainings are those performed in the physical Cyber Range. Also, to increase the number of training sessions, EDP also uses eLearning courses to sensitise employees for cybersecurity risks. The annual survey performed is focused on evaluating the global satisfaction with IT services in EDP, having a dedicated section to cybersecurity.

The training sessions in the Cyber Range are composed by the two parts, a theoretical introduction and a practice part. Basic concepts about cybersecurity are referred to during the theoretical introduction. By making the employees (trainees) interact with a cyber-attack scenario in the practice part, it is possible to see how they react, how aware they are of the need to report events, identify the root cause of the attack and define a strategy to tackle it. This training is accompanied and guided by trainers, allowing them to understand the level of awareness of trainees and adjust the training accordingly. All training sessions end with a debriefing stage for the revision of the cyber-attack, the analysis of the trainees' responses and receiving the feedback for the

improvement of the Cyber Range by using the Kirkpatrick Model [7]. The Kirkpatrick Model is a four-level training evaluation approach consisting of [7]:

- Level 1 (Reaction): The degree to which participants find the training favorable, engaging and relevant to their jobs.
- Level 2 (Learning): The degree to which participants acquire the intended knowledge, skills, attitude, confidence and commitment based on their participation in the training.
- Level 3 (Behaviour): The degree to which participants apply what they learned during training when they are back on the job.
- Level 4 (Results): The degree to which targeted outcomes occur as a result of the training.

3.1.4 Current evaluation practices by Ferrovie dello Stato Italiane

With the general objective of increasing awareness and risk culture throughout the Group, Ferrovie dello Stato Italiane, in collaboration with the Italian Postal Police, carried out a first training program aimed to show and demonstrate to employees the risks to which they are exposed in their daily use of computer devices.

The course consists of a first "teaser" dedicated to the presentation of the training plan, and four video learning pills lasting about 3 minutes. Each video, based on actors interpreting a short fiction, goes through the simulation of a real cyber-risk case. The video learning pills show the most common dangers and how to avoid them.

The contents of the video pills are reported below:

- Teaser: risks in sharing business information
- First video pill: Do not transfer / share your company passwords
- Second video pill: Do not use company email addresses to create social profiles
- Third video pill: Do not share confidential information with commercial file sharing tools (e.g. we transfer)
- Fourth video pill: Do not send received e-mails to the company e-mail address on a private e-mail address.

The course also encompasses an interview with Ivano Gabrielli, head of CNAIP (National IT crime protection centre for critical infrastructure).

After performing the course and watching all the video pills, a final learning test is proposed to the employee. The test recalls the topics of the fiction and proposes 10 multiple-choice questions in which only one answer is correct. The right answer is linked with Company Policies and Guidelines on the subject of Cyber Security.

3.2 What performance evaluation relevant information can be captured by the CYBERWISER.eu platform?

The CYBERWISER.eu Platform monitors the activities in the simulated environment with several components at various points in the infrastructure, on different levels and from different sources (sensors, scripts triggered by the platform at scheduled points in time, trainees' answers, etc.). Information that can be derived from the data coming from these sources which can be relevant for performance evaluation is described below, according to the components providing the information.

3.2.1 Economic Risk Evaluator

The Economic Risk Evaluator (ERE) tracks in real-time the cyber risk exposure of the monitored simulated platform. The measurement of the risk is done in economic terms. This is relevant as it may help infer how the blue and red teams are performing, with respect to risk exposure. Detailed information about the Economic Risk Evaluator is provided in Deliverable D2.5.

The Economic Risk Evaluator assesses the risk considering four different sources of information:

- A configuration questionnaire that serves to profile the organization for which the scenario is to be simulated. Different aspects are addressed, as they impact the risk profile, like the location of the company, the vertical it belongs to, the size, if they use cloud services, if they follow a BYOD (Bring Your Own Device) policy, IT security governance in place, how privacy is managed, how data are protected or past cyber risk episodes, to name but some.
- Target configuration based on the scenario definition, basically the IP/port of each element to include in the exercise ("element" meaning machine / application) and how important they are in terms of confidentiality, integrity and availability of the data they host. This importance is measured in a 0-10 scale. This information is needed to calculate the risk in terms of typical and worst economic losses.
- Events coming from the monitoring sensors and eventually processed by the Anomaly Detection Reasoner (see Section 3.2.2) to raise relevant alerts. This informs in real time of ongoing anomalies that may impact the cyber-risk exposure related to the monitored infrastructure.
- Vulnerabilities detected in the simulated infrastructure that may be exploited by a black-hat hacker.

The different risk models that will be produced during the CYBERWISER.eu project, both those that come from WISER and will be adapted and upgraded, and those that are created from scratch will provide means for the evaluation of a wide landscape of risks. This catalogue of models will be maintained and updated in connection with the existing trends in terms of threats and attacks both in Europe and globally.

Depending on the type of exercise that will be running, the measure of certain risks will be relevant. As stated below, the ERE follows a clear methodology to be able to inform about a wide catalogue of risks. Some examples are the following: SQL injection, cross-site scripting (XSS), session hijacking, phishing, password cracking, denial of service (DoS), bypass login by brute force, DNS login attack, attacks with Trojan, buffer overflow or client-server protocol manipulation, among others.

3.2.2 Monitoring Sensors and ADR

Monitoring sensors collect evidence from the monitored infrastructure, both at network and application levels, providing a first level of information aggregation without complex processing of the information. This evidence is interpreted by the ADR-Agent to match this information to likely types of events that may be taking place in the infrastructure. This information is sent to the ADR that filters and correlates it to obtain alarms. Unlike events, that can be considered only at warning level, the alarms require a reaction from the staff in charge of the security of the infrastructure. There may be different correlation levels, depending on the alarms that have been raised, which are linked to how much evidence there is about an ongoing incident. The higher the correlation level is, the more serious the alarm becomes and the more urgent the answer is. In short, the mission of the ADR is to distinguish what is abnormal in the infrastructure from what is normal.

The state of the art regarding sensor technology is rich in terms of variety of attacks that can be detected [8]. There are two types of efforts to be carried out in the project related to sensors: on the one hand the integration of them with the ADR-Agent and the ADR itself, and on the other hand we have to consider likely ad-hoc developments from scratch that may be necessary to satisfy needs posed by the exercises that are not covered by the current state of the art. Some examples of attacks that can be detected using a monitoring infrastructure composed by several sensors, the ADR-Agent and the ADR are: the denial of service both standard and distributed (DDoS), port scanning, brute force attack, SQL injection, trojans, USB detection, rootkits or fastflux attacks, to name but a few. Some sensors can inform about anomalies that can be indirectly linked to attacks such as a high number of connections or high/low network speeds. The presence of these attacks is a sign of good performance of the red team, while their short or long persistence might inform about the capacity of the blue team [11] [13] to react. In addition, the detection capabilities of the sensors can also be used to detect the application of mitigation measures by the user, which is an important element for evaluation: what mitigation was applied and when.

It is also relevant to mention that, as described in Section 3.2.1, the outputs of this monitoring infrastructure (events and alarms) are inputs used by the Economic Risk Evaluator, which offers an extra layer of intelligence to calculate the cyber risk exposure of the infrastructure.

3.2.3 Questionnaire

The advantage of the questionnaire is that allows both quantitative and qualitative metrics to be captured. This component can be implemented independently of the performance evaluator providing a simple but powerful tool for evaluation of the trainee in the offering levels where the performance evaluator is not available. Where the performance evaluator is available the questionnaire serves as another supplementary input allowing the platform to capture metrics not covered by the sensors.

The Questionnaire can cover both quantitative and qualitative metrics. Following a list of information that can be captured for each category. This is not intended as an exclusive list rather provides an indication of information that can be measured by this component to give a better understanding of the capabilities of the component.

- Quantitative metrics:
  - Access to sensitive information
  - Progress of an intrusion/training
  - Privilege escalation
  - Successful forensics analysis
  - Response time
- Qualitative metrics
  - Situational awareness
  - Decision reasoning
  - Analytical thinking

3.2.4 Countermeasure Simulator

The Countermeasure Simulator (CS) assists the blue team when they need support to apply certain mitigation measures in the context of a running exercise with some incidents taking place. The users might not need to use it if their expertise level is high, in such case the mitigations would be informed by means of monitoring sensors capabilities deployed. If they actually use the CS, this asset can produce logs informing about the application of the countermeasures. Apart from the information about what kind of measure is applied, there are some other important parameters that can be taken into account like the associated cost of the mitigation (to provide a more real user experience as mitigations usually do not come for free) or the rating, measured in terms of cost/benefit of the countermeasure.

These outputs become relevant input for the Performance Evaluator algorithms.

3.2.5 Attack Simulator

As detailed in Deliverable D2.5, Attack Simulator can be used by both white and red teams to launch attacks against the simulated infrastructure. In cyber-range exercises where the trainees' task is to defend the simulated infrastructure, Attack Simulator uses attack scripts provided by platform operators to automatically launch attacks according to pre-defined schedules adjustable by the white team [11][13]. Attack scripts output their success statuses that indicate whether an attack was successful, and this information is propagated in the form of an "exercise event" message to the Performance Evaluator. The criteria to consider an attack successful or not is very specific to the attack script and type of attack carried out. It can differ greatly from one attack to another, for example: a certain type of network connection was established to the target, some files were copied that should not be accessible, a database was dumped to the attacker, etc.

The data contained in such a message is:

- which attack script was launched,
- against which target in the simulated infrastructure (IP address),
- whether the attack was successful,
- time when the attack was launched.

Performance Evaluator can use this information to assess how successful the blue team was in defending their infrastructure. As cyber-range scenarios contain deliberately introduced vulnerabilities, the pre-defined attacks launched by the platform will succeed until the blue team implements the appropriate measures to prevent them. A failed attack towards a certain (initially vulnerable) target in the simulated environment indicates that the blue team was successful in mitigating the vulnerability.

When the Attack Simulator is used by the trainees, they can either use pre-defined attack scripts/templates or provide their own. Regardless, the scripts are open for manipulation by the trainees in this case and may not (properly) implement mechanisms to deduce success/failure of an attack, making them unsuitable for performance evaluation. Instead, to detect success of the attacks executed by the trainees (either via the Attack Simulator or launched from their own workstations), special custom sensors (see Section 3.2.2) can be used on the virtual machines simulating target infrastructure.

3.2.6 Vulnerability Assessment Tools

Vulnerability Assessment Tools (described in detail in Deliverable D2.5) offer two types of vulnerability scanners:

1) vulnerability scanners to detect generic vulnerabilities, and
2) detection scripts designed to detect other arbitrary vulnerabilities.

The result of a vulnerability scan in the first case is a report containing a list of vulnerabilities detected by the generic scanners. If a vulnerability featured in a specific cyber-range scenario can be detected by the scanners included in the generic suite, this report can be used to detect its presence and utilized for performance evaluation of blue teams, measuring their success in mitigation of vulnerabilities. Generic vulnerability reports are also used as input to the risk assessment procedure carried out by Economic Risk Evaluator.

If, on the other hand, a scenario contains a vulnerability that cannot be detected by the generic suite of scanners, specific vulnerability detection scripts can be used to report only the presence/absence of a particular vulnerability initially present in a scenario. Similar to the attack scripts used by the Attack Simulator, these scripts typically execute attacks towards the target and report their success. In contrast however, specific vulnerability checks execute passive attacks meaning that they generally don't change the state of the target machine.

Custom vulnerability detection scripts can also be used to detect downtime of services (periodically reporting whether a web service is available) or website defacement (reporting whether some content is present on a website).

Vulnerability scans of both types are triggered automatically with pre-defined schedules adjustable by the white team. The reports show whether the blue team was successful in mitigating certain vulnerabilities. A vulnerability scan report contains the data about:

- which scanner/vulnerability detection script was used,
- against which target in the simulated infrastructure (IP address),
- a list of detected vulnerabilities or a single Boolean value indicating the presence/absence of a specific vulnerability (depending on the type of scan),
- time of the scan.

3.2.7 Custom detectors of specific actions

Depending on the training scenario, trainees have specific objectives about tasks to complete on the simulated infrastructure. For example, the red team can be instructed to penetrate a target machine, obtain root access, and create a new user, introduce a channel for persistent (shell) access to the target, etc. The blue team might need to change some firewall settings to stop a certain type of network traffic. For the purpose of automatic performance evaluation, these actions can be detected using specifically implemented detectors.

As already described in Section 3.2.2, the sensors can be (depending on the required functionality) difficult to implement. Special consideration needs to be taken regarding isolation of such sensors to avoid trainees tampering with them.

These detectors are not parts of other CYBERWISER.eu components and will be implemented separately if needed and according to the needs of a particular training scenario. They will be implemented as simple programs that monitor a specific action or state on the target machine and send an appropriate message when a change is detected. Examples of actions that can be detected include:

- A root shell is opened on the machine: checking if a sh/bash process is being run by user root
- A new user was created on the machine: checking changes in the list of users
- A specific type of network traffic is being received: listening and filtering traffic using tcpdump and reporting on new packets received / flows established
- A specific file was changed on the machine: monitoring modified time / file hash / specific file contents
- etc.

Some of such actions can also be detected by using the monitoring sensors already available for deployment on the target machines (e.g. OSSEC). In this case, a custom rule needs to be set up for the Performance Evaluator to understand the sensor's event message as information, important for the performance evaluation.

3.3 State of the art

This section describes the relevant state of the art in terms of the evaluation of participants in context of courses. This includes: Bloom's taxonomy, Goal-Question-Metric, privacy scorecards, VALUE rubrics, and a Reflective Approach to Assessing Student Performance in Cyber-security Exercises [6].

3.3.1 Bloom's Taxonomy

The goal of writing the learning objectives for courses in an effective way has been achieved by using the Bloom's Taxonomy, which is actively used in Task T4.1 as part of developing courses. The Bloom's Taxonomy is a widely used and well known classification of the different objectives and skills, namely learning objectives, that educators set for their trainees (course participants) [1], [9]. The Taxonomy was proposed in 1956 by Benjamin Bloom and has been updated over the years to include the following six levels of learning [1]:

1. Remembering;
2. Understanding;
3. Applying;
4. Analysing;
5. Evaluating;
6. Creating.

Like other similar taxonomies, the Bloom's taxonomy is hierarchical, where *Remembering* is the lowest level and *Creating* is the highest level. This means that learning at the higher levels is dependent on having prerequisite knowledge and skills obtained at the lower levels. However, it is not mandatory to start with the lowest level and step all the way through the entire taxonomy. The Consortium found that the Bloom's taxonomy is a powerful tool to help develop learning objectives because it explains the process of learning. Indeed, before you can understand a concept, you must remember it. To apply a concept, you must first

understand it. In order to evaluate a process, you must have analysed it. To create an accurate conclusion, you must perform an overall evaluation.

As illustrated in Figure 3, Bloom's Taxonomy comes with a set of "action verbs" to help identifying appropriate exercises aligned with each level. The taxonomy is illustrated in Figure 3. One may notice that some of the verbs may be associated with multiple levels although they appear in separate levels. As an example, it is possible to have an objective stating "at the end of the lesson, trainees will be able to explain the difference between an XSS and a CSFR attack". This example would be an objective at the *understanding* level. However, the same verb can be used in an objective stating "a trainee can explain how a Wormhole attack can lead to a DoS". In this case, we have an objective that fits better in the *analysing* level. To write effective learning objectives there are few rules to follow. It is important to ensure the presence of one measurable verb in each objective. Each objective needs to include one action verb. It is important that verbs in the course level objective are at least at the higher Bloom's Taxonomy as the highest lesson level that support it. For example, it is impossible for trainees to *evaluate* a method/process if the lesson is only teaching how to *apply* the method/process. When writing course objectives, they must be written such that they are measurable, clear and concise.

The resulting learning objectives, obtained by applying the Bloom's Taxonomy, give some hints about the criteria which need to be considered in the process of evaluating the student's performance. One of the points stated above is the importance of using measurable verbs when writing learning objectives. The meaning behind this is that verbs which cannot be quantified express concepts difficult to measure, in terms of trainees' evaluation. As an example, verbs like understand, learn, appreciate and enjoy are barely assessable. Another important hint coming from learning goals written using the Bloom's Taxonomy, is that trainees' evaluation needs to be aligned with learning objectives. Verbs used in the objectives are helpful to identify appropriate exercises to measure the degree to which the learning objective is achieved. As an example, if a course has a learning objective using the action verb "explain" (*Understanding* level), we may create exercises asking the participant to draw/illustrate a diagram to explain, write a report, or answer a multiple-choice test, to mention a few examples [10].

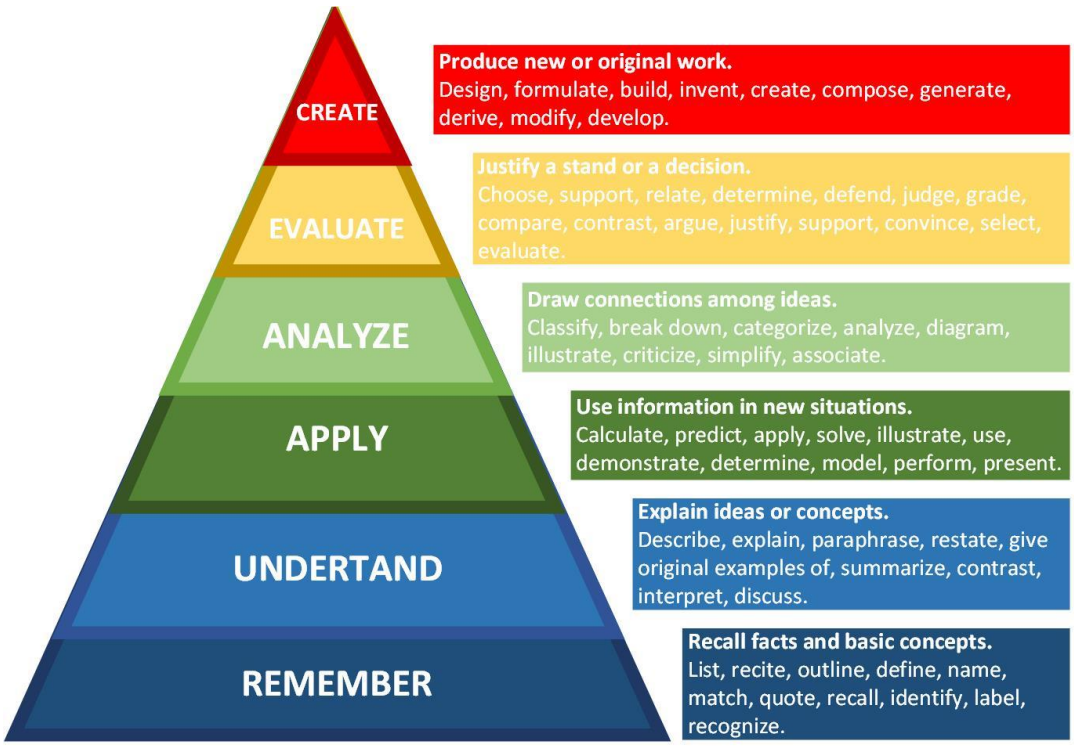

Figure 3. Bloom's Taxonomy and related action verbs

3.3.2 Goal-Question-Metric

The Goal-Question-Metric approach (GQM) stems from the desire to evaluate an organization or a project in a purposeful way, originally defined to evaluate defects for a set of NASA projects. The approach was originally intended to define and evaluate goals for specific projects but have evolved into a broader practice of quality improvement, especially within the software development industry. In general terms, GQM results in a specified measurement system focused on a set of cases, with a set of rules to interpret the measurement data [4]. Several earlier studies indicate that measurement need to be defined top-down, based on goals and models, hence the following hierarchical structure of GQM:

- Conceptual level. A goal is defined with respect to various models of quality.
- Operational level. A set of questions is framed to specify achievement of a particular goal.
- Quantitative level. A data set is connected to every question in order to give quantitative answers.

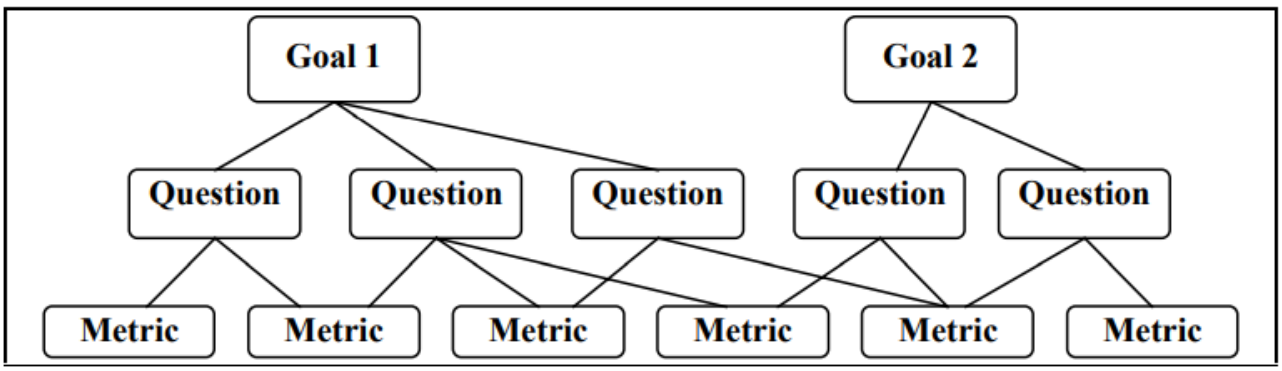

Figure 4. Goal-Question-Metric hierarchical structure

GQM can be undertaken in six-step process, where the first three target identifying the right metrics based on defined goals, and the last three target measurement gathering and effective usage of results to drive improvement.

1. Develop goals or objectives.
2. Generate questions to quantitatively define the developed goals.
3. Specify measurements needed to answer the generated questions.
4. Develop mechanisms to collect intended data.
5. Collect, validate and analyse data in real time.
6. Assess conformance to developed goals and make recommendations for improvement.

The key feature to incorporate, in our case, is to start by developing learning objectives (goals) for the trainees. Furthermore, GQM will assist in identifying indicators based on the defined learning objectives, and then to specify metrics needed to quantify or determine these indicators, and how to collect the required metrics. With quantitative answers to questions derived by the learning objectives, we can develop a mathematical function to generate a numerical performance evaluation – both for the individual objectives and for an overall performance.

3.3.3 Privacy Scorecard

To meet the increased service complexity and demand, services increasingly depend on data traceable to an individual, termed "personally identifiable information" (PII) [2] The attention and handling of PII through the intricate service ecosystem, diverges considerably between service providers. Moreover, insight and influence into treatment of PII among end-users diverges considerably as well. In order to maintain user trust regarding privacy, and to enforce the technical and legal requirements, Omerovic et Al. presents the "Privacy Scorecard" as a practical decision support for service providers [2] This approach has been shown to identify and specify privacy-relevant concerns of the service, and to gain new knowledge about the design of a service. For the consortium, it will serve as a basis for structuring the overall criteria for performance evaluation. The Privacy Scorecard is structured in a manner where the leftmost column is the top abstraction level, and every column following to the right is related to the one on its left.

Main concern	Success criteria	Indicators	Current score	Target score	Initiatives
Information to the user	A complete, comprehensible and correct assessment of privacy is presented to the user.	1. User consent is decomposed into relevant topics 2. Percentage of the relevant topics covered in the consent 3. Contents of the consent can easily be modified 4. Consent is easy to understand 5. The information in the content is up-to-date 6. The information in the consent is sufficient 7. Average time taken to update the information to the user, after a privacy-relevant change of the service	1. No 2. 5 of 10 3. No 4. No 5. 8 of 10 6. 9 of 10 7. 3 hours	1. Yes 2. 8 3. Yes 4. Yes 5. 8 6. 9 7. 3 hours	1. Improve usability of consent 2. Map needs for consent updates upon service upgrade/change 3. Inform the user about privacy risks
Retrieval and storing of PII	The service provider has access to correct, sufficient and only necessary PII. The PII is handled in a secure manner.	1. All PII has a valid purpose 2. PII is updated upon changes in the authoritative system that the data originates from 3. Number of unwanted incidents involving PII 4. There exist procedures for automatic deletion of PII			
Usage of PII	PII is used only for the purpose agreed between the service provider and the user.	PII is not combined without both agreement with user and a valid purpose			
Exchange of PII to a third party	PII is exchanged with a third party only if and when agreed upon with user, and the exchange is conducted in a secure manner. The third party is obliged to handle the PII in a secure manner.	1. PII is only exchanged upon informed consent from the user 2. PII is exchanged over an encrypted communication line			
User's control over own PII	The user can access and see all own PII which is stored and used by the service. The used can request all own PII to be removed. The user can request PII to fully or partially be transferred to other services.	1. The user has access to all own PII 2. There exist procedures for deletion of PII upon request of a user 3. Time needed to delete a PII, upon request from a user 4. Time needed to port/transfer PII, upon request from a user			

Figure 5. The initial generic Privacy Scorecard proposed by Omerovic et al. [2].

A main concern is depicted through success criteria, the success criteria is explored into a set of quantifiable indicators, a given indicator score is measured to a target score and lastly some initiatives may be outlined with regards to indicator shortcomings. Omerovic et Al. advocate the following guidance for using the Privacy Scorecard [2]:

1. Specify the target of the analysis
2. Identify the privacy regulations, requirements, strategies and goals of the commissioner with respect to the system/service of the analysis.
3. Identify main privacy concerns.
4. Explain meaning of each concern through success criteria in column two.
5. Identify and specify indicators relevant to each concern.
6. Specify the target score of each indicator.
7. Identify and specify the initiatives.
8. Specify revision plans.

The approach of generating "Privacy Scorecard" relates to that of GQM, as does the structure of the scorecard's three leftmost column to the GQM hierarchy. The "Privacy Scorecard" template can provide itself useful as a basis for the performance evaluation, where learning objectives relate to concerns, success criteria relate to evaluation criteria and indicators become quantifiable.

3.3.4 VALUE Rubrics

During the period of 2007-2009 the VALUE [14] (Valid Assessment of Learning in Undergraduate Education) initiative gathered faculty and educational professionals from over 100 higher educational institutions, to develop VALUE rubrics for the LEAP [15] (Liberal Education and America's Promise) Essential Learning Outcomes. These rubrics seek to cover every level in Bloom's taxonomy, and every aspect in applying knowledge in the real world. This includes aspects such as integrative and applied learning, personal and social responsibility, intellectual and practical skills, ethical reasoning, teamwork etc., which all are integral to the realm of cybersecurity. For each learning outcome, characteristics or criteria were identified, and the most frequent lay the groundwork the respective learning outcome [5]. Among the Essential Learning Outcomes, the following adjures for mentioning by relevance to this consortium:

- Inquiry and analysis,
- Problem solving,
- Integrative learning.

Each rubric articulates fundamental criteria for each learning outcome, with performance description and four corresponding evaluation levels of learning. The problem-solving VALUE rubric, of Table 3, identifies the following criteria and evaluator descriptions.

	Capstone 4	**Milestone 3**	**Milestone 2**	**Benchmark 1**
Define Problem	Demonstrates the ability to construct a clear and insightful problem statement with evidence of all relevant contextual factors.	Demonstrates the ability to construct a problem statement with evidence of most relevant contextual factors, and problem statement is adequately detailed.	Begins to demonstrate the ability to construct a problem statement with evidence of most relevant contextual factors, but problem statement is superficial.	Demonstrates a limited ability in identifying a problem statement or related contextual factors.
Identify Strategies	Identifies multiple approaches for solving the problem that apply within a specific context.	Identifies multiple approaches for solving the problem, only some of which apply within a specific context.	Identifies only a single approach for solving the problem that does apply within a specific context.	Identifies one or more approaches for solving the problem that do not apply within a specific context.
Propose Solutions	Proposes one or more	Proposes one or more	Proposes one solution/hypothesis	Proposes a solution/hypothesis

	Capstone 4	Milestone 3	Milestone 2	Benchmark 1
	solutions/hypotheses that indicates a deep comprehension of the problem. Solution/hypotheses are sensitive to contextual factors (ethical and logical dimensions).	solutions that indicates comprehension of the problem. Solutions are sensitive to contextual factors (ethical and logical dimensions)	that is “off the shelf ” rather than individually designed to address the specific contextual factors of the problem.	that is difficult to evaluate because it is vague or only indirectly addresses the problem statement.
Evaluate Potential Solutions	Evaluation of solutions is deep and elegant (for example, contains thorough and insightful explanation) and includes, deeply and thoroughly, all of the following: considers history of problem, reviews logic/reasoning, examines feasibility of solution, and weighs impacts or solutions.	Evaluation of solutions is adequate and includes the following: considers history of problem, reviews logic/reasoning, examines feasibility of solution, and weighs impacts of solution.	Evaluation of solutions is brief (for example, explanation lacks depth) and includes the following: considers history of problem, reviews logic/reasoning, examines feasibility of solution, and weighs impacts of solution	Evaluation of solutions is superficial (for example, contains cursory, surface level explanation) and includes the following: considers history of problem, reviews logic/reasoning, examines feasibility of solution, and weighs impacts of solution.
Implement Solutions	Implements the solution in a manner that addresses thoroughly and deeply multiple contextual factors of the problem.	Implements the solution in a manner that addresses multiple contextual factors of the problem in a surface manner.	Implements the solution in a manner that addresses the problem statement but ignores relevant contextual factors.	Implements the solution in a manner that does not directly address the problem statement.
Evaluate Outcomes	Reviews results relative to the problem defined with thorough, specific considerations of need for further work.	Reviews results relative to the problem defined with some consideration of need for further work.	Reviews results in terms of the problem defined with little, if any, consideration of need for further work	Reviews results superficially in terms of the problem defined with no consideration of need for further work

Table 3. Problem solving value rubric

When defining thresholds and levels of achievement, e.g. to enable translation from a numerical evaluation criterion into descriptive performance level and feedback, the abovementioned rubric would be a useful tool for the consortium. The four defined levels of learning could be a basis for categorization of performance, and the indicators for each evaluation criterion could be mapped into one of the essential criteria for the learning outcome. This VALUE rubric also serves to highlight actual objectives of learning, which is severe to being able to evaluate a trainee's level of understanding through a particular - or a set of - exercises.

3.3.5 A Reflective Approach to Assessing Student Performance in Cyber-security Exercises

If on targets to assess student performance, in a cyber-security, on a thorough level, evaluation needs to be extended from only quantifying some predefined indicators, such as time, complete or incomplete, correct or incorrect etc. When evaluating a student's skill level in solving a cyber-security exercise, the train of thought could provide an even deeper insight. In A Reflective Approach to Assessing Student Performance in Cybersecurity Exercises, Weiss et al. explore the use of command line history in order to obtain additional information about the students' solution strategy when solving a cybersecurity exercise [6]. The exercise studied consists of determining the number of live hosts (if any) along with any running network services on those hosts in a network, within a timeframe of one hour. By capturing the Linux bash history of commands entered by the students, they desired to gain insight into and quantify the "solution path". The commands were graphed to visualize the "solution path", and then the trainers performed a qualitative analysis of the given path, based on their knowledge, path length, path extensiveness and quantitative outcome. From studying seven teams undertaking the exercise, they concluded that a simple indicator relying on identifying the correct IP addresses would not have been an accurate assessment of their understanding, as some achieved the results by chance rather than knowledge. Weiss et al. outline some experiences and recommendations [6]:

- A graph-based trace of students' solution strategies can constitute valuable feedback to the students.
- It can be used as a review artefact. Each trace can be used to identify pros and cons of the students' solution approach.
- It can be used in an automated "live" scoring method for competitions. Capture-the-flag style competitions as popular cybersecurity exercises, and live performance can often benefit both the trainees and the instructors.

For this consortium, a way of identifying the trainee's approach to solving the exercise would enhance the performance evaluation greatly, according to Weiss et al. This could comprise meaningful insight into the development of more complex indicators, that now only are directly measurable metrics (e.g. time) but are obtained through an algorithmic analysis of students' solutions strategies.

4. Guidelines for identifying and specifying criteria for the evaluation of trainee performance

This section describes the overall method used to identify and specify evaluation criteria for the purpose of evaluating trainees as part of the cyber training courses (developed in Task T4.1) and exercises (developed in Task T4.2). To support the steps of the method, we defined templates to specify evaluation criteria as well as indicators used to assess how well an evaluation criterion has been achieved. Section 4.1 describes the method, while Section 4.2 describes the template for indicators and the template for evaluation criterion used in the method.

The process of identifying evaluation criteria is carried out in relation to defining learning goals and objectives as well as the teaching and learning activities [9]. Figure 6 illustrates this relationship. According to Fink [9], learning goals are described at high level, while objectives are refinements of goals providing more details. These are commonly defined as part of course descriptions. In the teaching and learning activities we are concerned about developing activities to make sure that trainees have learned the knowledge required to achieve the goals and objectives. Finally, within feedback and assessment, we define evaluation criteria to assess if trainees have achieved the goals and objectives.

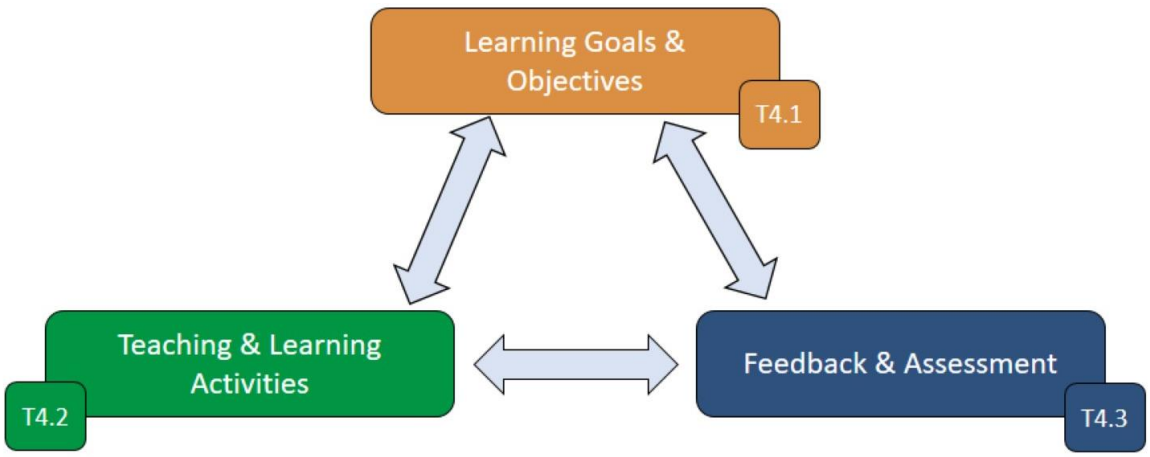

Figure 6. The relationship between learning goals & objectives, teaching & learning activities, and feedback & assessment (adapted from Fink [9])

Figure 6 also illustrates (roughly) the tasks in CYBERWISER.eu that mainly address each aspect. The learning goals and objectives are developed as part of course descriptions in Task T4.1 which also contains a brief description of associated teaching and learning activities. Although the course descriptions also contain a brief description of the teaching and learning activities, these are developed in detail in Task T4.2. The definition of appropriate evaluation criteria is carried out in Task T4.3 and documented in this report.

The evaluation criteria defined in Section 5 have been developed considering the learning goals and objectives of the exercises for the pilots in WP5 (defined in Task T5.1 and documented in Deliverable D5.1) and the expected corresponding courses to be developed in Task T4.1 for the intermediate and advanced offering levels. There are two main reasons to this:

1. The pilots in WP5 are interested in certain learning goals and objectives that require the availability of the technical assets: Performance Evaluator, Economic Risk Evaluator, Countermeasure Simulator, and Vulnerability Assessment Tools.
2. According to the CYBERWISER.eu learning path defined in Deliverable D4.1 [12], courses in which the above technical assets are used will be available in the Intermediate and Advanced offering levels.

In this respect, the evaluation criteria defined in Section 5 requires all technical assets to support real-time performance evaluation. However, the list of evaluation criteria will be extended as new courses will be defined in Task T4.1, as well as considering potential new learning goals and objectives defined by the pilots in WP5. This extension will be carried out as part of the validation activities of CYBERWISER.eu.

4.1 Method for identifying and specifying evaluation criteria

Figure 7 illustrates the four steps of the method to identify and specify the evaluation criteria. The first two steps are to identify learning goals and learning objectives, respectively. This is carried out as part of course descriptions. The reader is referred to Deliverable D4.1 [12] for the process of identifying and describing courses.

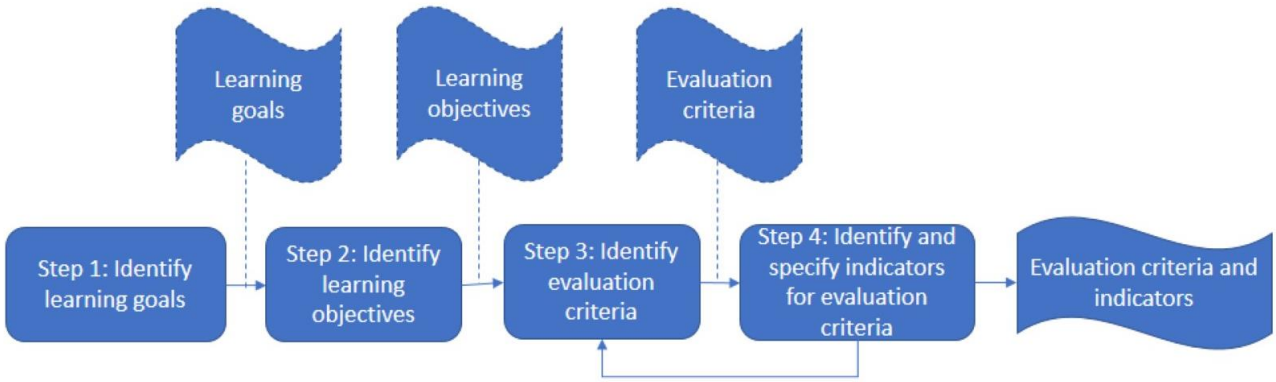

Figure 7. Method for identifying and specifying evaluation criteria

Before continuing on to the next steps, it is important to note that one learning objective must have at least one associated exercise (teaching & learning activities in Figure 6) because the exercises are where the evaluation criteria are applied in order to evaluate the performance of a trainee. Task 4.2 is at the time of writing developing a cyber-training scenario development method, which will be reported in deliverables D4.3 and D4.5, and this method can be used to develop appropriate exercises for a given learning objective.

Having defined learning goals and learning objectives, the third step is to identify evaluation criteria. An evaluation criterion is defined using the evaluation criterion template in Table 5. Notice that the evaluation criterion template considers an objective as well as an exercise related to the objective. Finally, having defined a set of evaluation criteria, we identify a set of indicators for each evaluation criterion. An indicator is defined using the indicator template in Table 4. The indicators are used collectively to assess the achievement of an evaluation criterion. The indicator template in Table 4 considers also the frequency of measurement. The measurement frequency of the evaluation criteria and their supporting indicators defined in Section 5 are defined to facilitate dynamic performance evaluation of trainees.

Sometimes it may be difficult to first identify an evaluation criterion and then a set of indicators for that evaluation criterion. In such situations, based on our experience, we find it is easier to start by defining a set of indicators and then define an evaluation criterion considering the identified indicators (thus, the arrow going back from Step 4 to Step 3).

The result of the method described is a set of evaluation criteria and associated indicators.

4.2 Templates for specification of the criteria

The templates applied in the method described in Section 4.1 are described in detail in this section. Indicators represent concrete measures obtainable via the automatic monitoring capabilities in CYBERWISER.eu to support real-time evaluation of the performance of a trainee in terms of problem solving in the exercises. For example, the time a trainee uses to carry out one specific exercise. An indicator may also be described as non-automatic in the sense that a human instructor (e.g., the trainer) can assess a value. For example, assessing the analytical skills of a trainee may require that a trainer assesses the analytical skill based on expert knowledge and know-how, which is most likely not possible to obtain via automatic monitoring capabilities.

Indicators are used collectively in an evaluation criterion to assess the degree to which the evaluation criterion has been achieved. This is done by using the indicator values "as is". By reporting the level of achievement using indicator values "as is" provides a precise assessment of "how well" a trainee has achieved the evaluation criterion. The template used to describe indicators is provided in Table 4. The indicator template in Table 4 has three main columns. The attribute column provides the "meta" information about an indicator, the description column provides a description for each attribute to guide the user in defining the indicator, while the mandatory column indicates which attributes are mandatory when describing an indicator. When filling out the indicator template, the user removes the guiding text in all rows of the "description of the attribute" column and writes appropriate text according to the guidelines. The following assumptions were made when developing the indicator template:

- An indicator may be exercise specific or exercise agnostic.
- An indicator needs to be measurable.
- An indicator needs to be an underlying input to an evaluation criterion to support the assessment of how well an evaluation criterion has been achieved.
- If uncertainty exists with regards to the measurement of an indicator, this needs to be explicitly outlined under the "Uncertainty" attribute.

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Unique ID of the indicator (e.g. pilot.indicator_number)	X
Name	A short name of the indicator	X
Definition	Definition of the indicator, and the variables/parameters. (e.g. a flag, or time)	X
Purpose	What purpose the indicator serves?	X
Measurement procedure	Specifies how to retrieve the indicator values.	
Data source	Specifies where to retrieve the indicator values.	
Measurement frequency	Specifies how often to retrieve the indicator values (or what triggers a new measurement?)	X
Expected change frequency	Specifies how often the indicator values are expected to change in reality (i.e. the dynamics of the indicator)	X
Unit of measure	Specifies the unit of measure of the indicator.	
Interpretation of the value measured	Specifies ranges of the indicator values, e.g. desirable, realistic but extreme, common, the edge to the acceptable, the edge to the unacceptable.	X

Attribute	Description of the attribute	Mandatory
Scale	Specified the measurement scale for the indicator.	
Uncertainty	Specifies the uncertainty and the related sources to it. Can be expressed in the form of interval, variance, comments, etc.	X
Storage	Specifies under which constraints the indicator is valid and for how long it should be stored.	
Value, exercise, user, and measurement date	Actual indicator value, the exercise it is linked to, user (trainee/trainer) it is related to, and the date of value retrieval.	

Table 4. Indicator template (based on Omerovic et al. [2], [3]).

The evaluation criteria are used to assess how well a trainee has achieved a learning objective. As mentioned in Section 3.3.1, an example of a learning objective is "at the end of the lesson, trainees will be able to explain the difference between an XSS and a CSFR attack". This learning objective may be decomposed into a set of evaluation criteria which can help assessing whether the learning objective has been achieved and how well it has been achieved. The reader is referred to Deliverable D4.1 [12] for a detailed explanation of learning objectives and those defined for the courses provided in Primer and Basic offering levels. The template used to describe an evaluation criterion is provided in Table 5. The user fills out the template in a similar manner as for the indicator template described above. The following assumptions were made when developing the evaluation criterion template:

- If uncertainty exists with regards to ambiguity around level of criterion fulfilment, this needs to be explicitly outlined under the "Uncertainty" attribute.
- "Interpretation of the score obtained" can require human (trainer) interpretation.
- There are one or more evaluation criteria per exercise. A criterion is exercise specific.
- A criterion is assessed with respect to its supporting indicators.
- A criterion needs to be derived from an educational objective.

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Unique ID of the evaluation criterion	X
Name	A short name of the evaluation criterion	X
Exercise	Which specific exercise the criterion is related to.	X
Educational objectives	The educational objectives of the specific exercise the criterion is related to.	
Underlying indicators	IDs of the indicators relevant for this evaluation criterion.	X
Aggregation	Defines the aggregation procedure from indicators to the evaluation criterion of the exercise in question, in order to evaluate performance of a trainee within this specific exercise.	X
Update frequency	Specifies how often to update the evaluation criterion score (or what triggers a new update). E.g. once per exercise execution.	X

Attribute	Description of the attribute	Mandatory
Interpretation of the score obtained	Specifies ranges of the evaluation criterion scores, e.g. desirable, realistic but extreme, common, the edge to the acceptable, the edge to the unacceptable.	X
Scale	Specified the measurement scale for the evaluation criterion.	
Uncertainty	Specifies the uncertainty and the related sources to it. Can be expressed in the form of interval, variance, comments, etc.	X
Storage	Specifies under which constraints the evaluation criterion is valid and for how long it should be stored.	
Value, exercise, user, and measurement date	Actual score of the evaluation criterion, user (trainee/trainer) it is related to, and the date of score retrieval.	

Table 5. Evaluation criterion template (based on Omerovic et al. [2], [3]).

5. The established evaluation criteria and indicators

This section describes the identified evaluation criteria as reported by the pilots in WP5 using the method and the templates described in Section 4. As mentioned in Section 4, the evaluation criteria have been developed by considering the learning goals and objectives of the exercises for the pilots in WP5 (defined in Task T5.1) and the expected corresponding courses to be developed in Task T4.1 for the intermediate and advanced offering levels. This section is organized as follows:

- Section 5.1 describes the evaluation criteria and supporting indicators relevant for Pilot 1, Academic training.
- Section 5.2 describes the evaluation criteria and supporting indicators relevant for Pilot 2, Transport infrastructure.
- Section 5.3 describes the evaluation criteria and supporting indicators relevant for Pilot 3, Energy infrastructure.

Each of the abovementioned sections provides first an overview of the identified evaluation criteria and the indicators supporting those evaluation criteria. Then, the evaluation criteria and indicators are described in detail according to the templates in Section 4.2. Each evaluation criterion description contains a reference to one relevant exercise described in Deliverable D5.1.

5.1 Evaluation criteria and indicators for Pilot 1 (academic training)

Figure 8 provides an overview of the evaluation criteria and their supporting indicators identified for Pilot 1, Academic Training.

Figure 8 illustrates also the relationship between the evaluation criteria and the indicators. For example, the evaluation criterion ATC1 Network traffic blocking (described in Table 8) uses the indicators ATI1 Time (described in Table 6) and ATI2 Traffic blocked (described in Table 7). ATC is short for Academic Training Criterion, while ATI is short for Academic Training Indicator.

As illustrated in Figure 8, there are in total 8 evaluation criteria (ATC1 – ATC8), and in total 12 indicators (ATI1 – ATI12) identified for Pilot 1, Academic training.

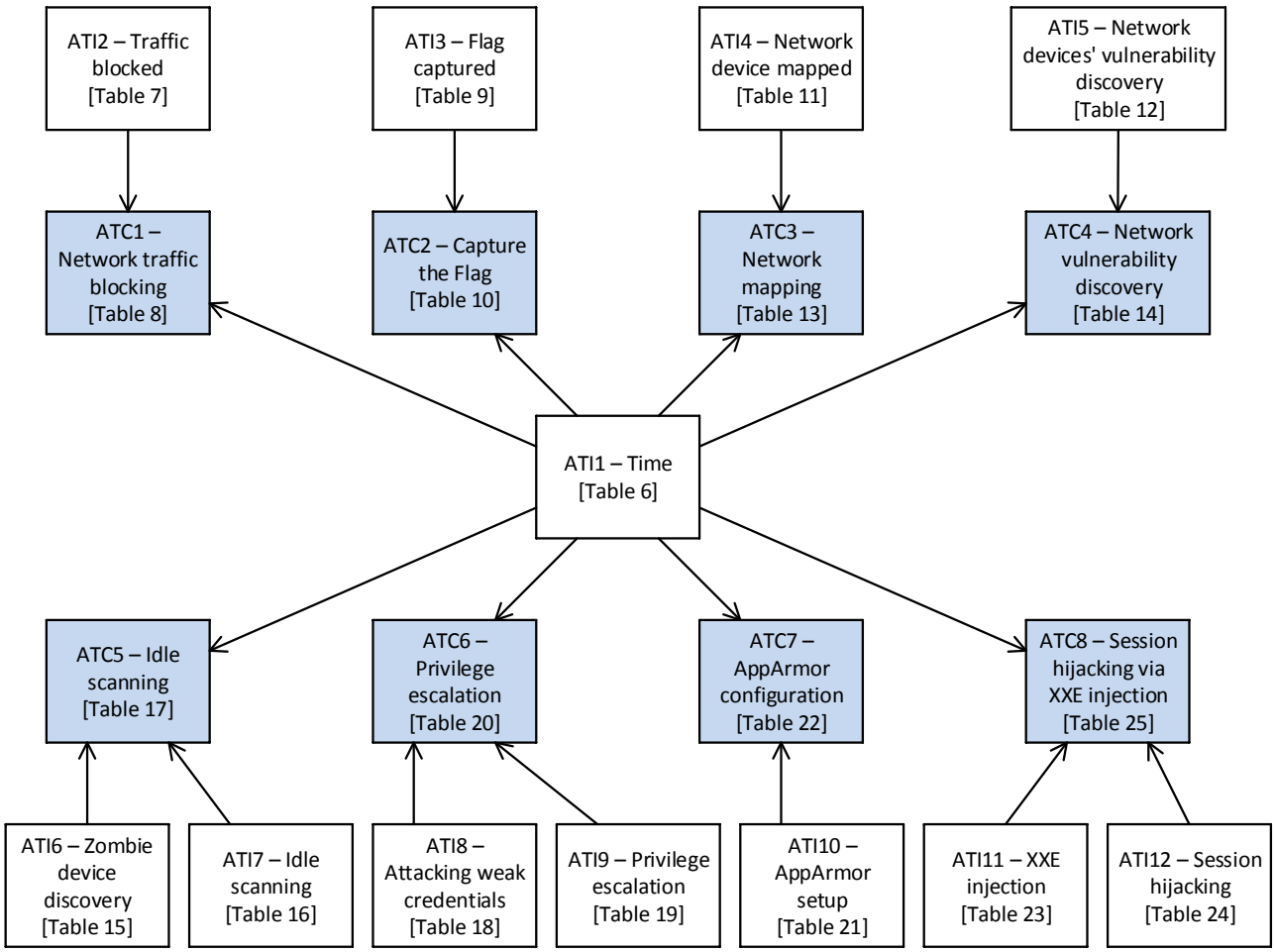

Figure 8. Overview of evaluation criteria and supporting indicators for Pilot 1, Academic Training

The following tables describe the evaluation criteria (Table 8, Table 10, Table 13, Table 14, Table 17, Table 20, Table 22, Table 25) as well as the indicators, as illustrated in Figure 8.

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_1	X
Name	Time	X
Definition	Maximum time to complete the exercise	X
Purpose	Limit the time duration of the exercise	X
Measurement procedure	The trainer will take the starting time when the exercise begins and define a known threshold for completing the exercise	
Data source	Trainer's clock system	
Measurement frequency	Twice per exercise: the measure is done at the beginning at the exercise, and at the end	X
Expected change frequency	By the end of the exercise, the time runs out	X

Attribute	Description of the attribute	Mandatory
Unit of measure	Seconds	
Interpretation of the value measured	Trainees shall complete the exercise within 2 hours, in order to complete correctly the exercise	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 6. Academic training indicator 1 - Time

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_2	X
Name	Traffic blocked	X
Definition	A particular kind of network traffic (e.g. network traffic over port 80) has been blocked	X
Purpose	Check that the correct firewall rule has been applied	X
Measurement procedure	The trainer will send malicious traffic to check that it is blocked and dropped by the firewall (e.g. by the use of netstat)	
Data source	Trainees' victim machine	
Measurement frequency	For each kind of traffic which trainees should block, one per exercise	X
Expected change frequency	After the trainer sends the malicious traffic	X
Unit of measure		
Interpretation of the value measured	If the traffic is blocked, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 7. Academic training indicator 2 - Traffic blocked

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Academic_criterion_1	X
Name	Network traffic blocking	X
Exercise	Firewall and network filtering (see D5.1 for detailed description)	X
Educational objectives	Creating new rules for filtering network traffic	
Underlying indicators	academic_indicator_1, academic_indicator_2	X
Aggregation	Trainees must be able to write effective rules in order to block malicious network traffic, within the time duration of the exercise	X
Update frequency	Once per exercise execution	X
Interpretation of the score obtained	If more than the 60% of the traffic kind has been blocked, within the specified time period, then the result is positive	X
Scale	The score is computed in terms of percentage of traffic kind filtered by the trainee in the specified time period	
Uncertainty	If the exercise requires to block more than one kind of traffic, an uncertainty is found on the case that trainees are able to block a certain percentage of malicious traffic. If it's the case, such percentage must be over 60% on order to correctly complete the exercise.	X
Storage		
Value, exercise, user, and measurement date		

Table 8. Academic training criterion 1 - Network traffic blocking

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_3	X
Name	Flag captured	X
Definition	A particular character string (flag) has been submitted by the trainee	X
Purpose	Check that a particular text find has been retrieved by the trainee	X
Measurement procedure	The submitted flag is compared to the correct one	
Data source		
Measurement frequency	Once per exercise	X

Attribute	Description of the attribute	Mandatory
Expected change frequency	Once per exercise, when the trainee submits the retrieved flag	X
Unit of measure		
Interpretation of the value measured	If the submitted flag is equal to the correct one, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 9. Academic training indicator 3 - Flag captured

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Academic_criterion_2	X
Name	Capture the Flag	X
Exercise	SQL Injection (see D5.1 for detailed description)	X
Educational objectives	Applying SQL statement to steal personal data	
Underlying indicators	academic_indicator_1, academic_indicator_3	X
Aggregation	Trainees must be able to apply the correct SQL statement in the vulnerable web form, within the time duration of the exercise	X
Update frequency	Once per exercise execution	X
Interpretation of the score obtained	If the correct flag is captured and submitted before the time runs out, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 10. Academic training criterion 2 - Capture the Flag

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_4	X
Name	Network device mapped	X
Definition	The description of the identified network device (IP and OS) is submitted by the trainee	X
Purpose	Check that the trainee maps the devices connected to the network	X
Measurement procedure	The submitted device's information are compared with the complete list of devices attached to the network, in order to check that it is present in the trainer's list	
Data source		
Measurement frequency	Once per exercise	X
Expected change frequency	Once per exercise, when the trainee submits the device's information	X
Unit of measure		
Interpretation of the value measured	If the submitted device's information is in the list of attached devices, then the result is positive	X
Scale		
Uncertainty	An uncertainty may happen if the submitted information are partially correct (wrong OS / correct IP and vice versa). In this case, the result is given by checking the IP	X
Storage		
Value, exercise, user, and measurement date		

Table 11. Academic training indicator 4 - Network device mapped

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_5	X
Name	Network devices' vulnerability discovery	X
Definition	The list of vulnerabilities for the devices attached to the network is submitted by the trainee	X
Purpose	Check that the trainee discovers the vulnerabilities for the devices connected to the network	X
Measurement procedure	The submitted device's vulnerability is compared with the complete list of devices' vulnerabilities attached to the network, in order to check that it is present in the trainer's list	

Attribute	Description of the attribute	Mandatory
Data source		
Measurement frequency	Once per exercise	X
Expected change frequency	Once per exercise, when the trainee submits the device's vulnerabilities	X
Unit of measure		
Interpretation of the value measured	If the submitted device's vulnerability is in the complete list of vulnerabilities, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 12. Academic training indicator 5 - Network devices' vulnerability discovery

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Academic_criterion_3	X
Name	Network mapping	X
Exercise	Network and vulnerability scanner (see D5.1 for detailed description)	X
Educational objectives	Illustrating the services provided by devices connected to the network by the mean of Nmap	
Underlying indicators	academic_indicator_1, academic_indicator_4	X
Aggregation	Trainees must be able to find the devices attached to the network within the time limits	X
Update frequency	Once per exercise execution	X
Interpretation of the score obtained	If the trainee is able to discover at least 60% of devices within the time limits, then the result is positive.	X
Scale	The score is computed in terms of percentage of devices' number within the time limits.	
Uncertainty	No uncertainty applicable to this criterion	X
Storage		
Value, exercise, user, and measurement date		

Table 13. Academic training criterion 3 - Network mapping

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Academic_criterion_4	X
Name	Network vulnerabilities discovery	X
Exercise	Network and vulnerability scanner (see D5.1 for detailed description)	X
Educational objectives	Evaluating if there are vulnerabilities on the network devices by the mean of OpenVas	
Underlying indicators	academic_indicator_1, academic_indicator_5	X
Aggregation	Trainees must be able to find the vulnerabilities of the devices attached to the network within the time limits	X
Update frequency	Once per exercise execution	X
Interpretation of the score obtained	If the trainee discovers at least 60% of devices' vulnerabilities within the time limit, then the result is positive.	X
Scale	The score is computed in terms of percentage of devices' vulnerabilities number within the time limits.	
Uncertainty	No uncertainty applicable to this criterion	X
Storage		
Value, exercise, user, and measurement date		

Table 14. Academic training criterion 4 - Network vulnerabilities discovery

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_6	X
Name	Zombie device discovery	X
Definition	Discover a device on the network which can be used as zombie	X
Purpose	Check that the trainee is able to use Nmap [16] for checking the possibility of using one of the network devices as zombie to launch an idle scan	X
Measurement procedure	The IP of the discovered zombie machine will be submitted and compared with the zombie machine identified by the trainer	
Data source		
Measurement frequency	Once per exercise	X
Expected change frequency	Once per exercise, when the trainee submits the IP of the zombie machine	X

Attribute	Description of the attribute	Mandatory
Unit of measure		
Interpretation of the value measured	If the IP submitted by the trainee is the same identified by the trainer, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 15. Academic training indicator 6 - Zombie device discovery

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_7	X
Name	Idle scanning	X
Definition	Launch an idle scan to discover the devices attached to the network	X
Purpose	Check that the trainee is able to use Nmap to launch an idle scan using a zombie device	X
Measurement procedure	The trainee will submit the identified network device in order to check that it is present in the trainer's list	
Data source		
Measurement frequency	Once per exercise	X
Expected change frequency	Once per exercise, when the trainee submits the identified device	X
Unit of measure		
Interpretation of the value measured	If the submitted network device is present in the trainer's list, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 16. Academic training indicator 7 - Idle scanning

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Academic_criterion_5	X
Name	Idle scanning	X
Exercise	Idle scan (see D5.1 for detailed description)	X
Educational objectives	Performing an idle scan in the LAN using Nmap [16]	
Underlying indicators	academic_indicator_1, academic_indicator_6, academic_indicator_7	X
Aggregation	Trainees must be able to list the devices attached to the network by the mean of a zombie machine, within the time limits	X
Update frequency	Once per exercise execution	X
Interpretation of the score obtained	If the trainee is able to discover at least 60% of devices attached to the network within the time limits, then the result is positive.	X
Scale	The score is computed in terms of percentage of devices number within the time limits.	
Uncertainty	No uncertainty applicable to this criterion	X
Storage		
Value, exercise, user, and measurement date		

Table 17. Academic training criterion 5 - Idle scanning

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_8	X
Name	Attacking weak credentials	X
Definition	Launch a dictionary attack to a Tomcat 8 web service [17]	X
Purpose	Check the skill of the trainee in launching a dictionary attack to retrieve weak credentials	X
Measurement procedure	The trainee will submit the identified username and password	
Data source		
Measurement frequency	Once per exercise	X
Expected change frequency	Once per exercise, when the trainee submits the credentials	X
Unit of measure		

Attribute	Description of the attribute	Mandatory
Interpretation of the value measured	If the submitted credentials are the same which the trainer set on the victim machine, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 18. Academic training indicator 8 - Attacking weak credentials

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_9	X
Name	Privilege escalation	X
Definition	Take advantage of a vulnerability in the Tomcat 8 [17] machine to perform a Privilege Escalation attack on the victim	X
Purpose	Check the ability of the trainee to launch an exploit on the victim machine in order to gain more privilege than the user with weak credentials	X
Measurement procedure	The trainee submits a secret text string which only a root user can read from the victim machine	
Data source		
Measurement frequency	Once per exercise	X
Expected change frequency	Once per exercise, when the trainee submits the text string	X
Unit of measure		
Interpretation of the value measured	If the submitted text string is the one prepared by the trainer, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 19. Academic training indicator 9 - Privilege escalation

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Academic_criterion_6	X
Name	Privilege escalation	X
Exercise	Privilege escalation (see D5.1 for detailed description)	X
Educational objectives	Demonstrating how the usage of weak credential can lead to unauthorized access.	
Underlying indicators	academic_indicator_1, academic_indicator_8, academic_indicator_9	X
Aggregation	The trainee has to perform a privilege escalation attack to retrieve a secret string file on a Tomcat [17] service's machine, after gaining access to the non-privileged user using a dictionary attack, within the time limits	X
Update frequency	Once per exercise execution	X
Interpretation of the score obtained	If the trainee is able to retrieve both weak credentials and the secret text string within the time limits, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this criterion	X
Storage		
Value, exercise, user, and measurement date		

Table 20. Academic training criterion 6 - Privilege escalation

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_10	X
Name	AppArmor setup	X
Definition	Correctly configure AppArmor [18] in order to keep the service running on the victim machine	X
Purpose	Check that the trainee has correctly configured AppArmor on the victim machine they are controlling, to avoid the service to be unavailable for legitimate user	X
Measurement procedure	Check that the service is available after the trainer launched the attack	
Data source		
Measurement frequency	Once per exercise	X
Expected change frequency	Once per exercise, after the trainer launched the attack	X

Attribute	Description of the attribute	Mandatory
Unit of measure		
Interpretation of the value measured	If the service is running after the trainer launched the attack, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 21. Academic training indicator 10 - AppArmor setup

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Academic_criterion_7	X
Name	AppArmor configuration	X
Exercise	AppArmor defence (see D5.1 for detailed description)	X
Educational objectives	Using AppArmor to defend a system from external intruders	
Underlying indicators	academic_indicator_1, academic_indicator_10	X
Aggregation	The trainee must configure AppArmor within a limited time to avoid external intrusions by a malicious client. Whether the trainee has successfully configured AppArmor to avoid intrusion attacks is tested by the Trainer launching the attack after the trainee has configured AppArmor within the limited time.	X
Update frequency	Once per exercise execution	X
Interpretation of the score obtained	If the trainee is able to retrieve both weak credentials and the secret text string within the time limits, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this criterion	X
Storage		
Value, exercise, user, and measurement date		

Table 22. Academic training criterion 7 - AppArmor configuration

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_11	X
Name	XXE injection	X
Definition	Perform an XXE injection on the victim machine	X
Purpose	Check trainee's ability in using BURP [19] to perform an XXE injection	X
Measurement procedure	The trainee will submit the stolen session id of another user	
Data source		
Measurement frequency	Once per exercise	X
Expected change frequency	Once per exercise, after the trainee submitted the session id	X
Unit of measure		
Interpretation of the value measured	If the session id is equal to the one supplied by the trainer, the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 23. Academic training indicator 11 - XXE injection

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	academic_indicator_12	X
Name	Session Hijacking	X
Definition	Login in the victim machine using a session id of another user to retrieve a secret text string	X
Purpose	Check trainee's ability in using BURP [19] to perform an XXE injection	X
Measurement procedure	The trainee will submit the secret text string retrieved by the victim machine	
Data source		
Measurement frequency	Once per exercise	X

Attribute	Description of the attribute	Mandatory
Expected change frequency	Once per exercise, after the trainee submitted the secret text string	X
Unit of measure		
Interpretation of the value measured	If the secret text string is equal to the one supplied by the trainer, the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 24. Academic training indicator 12 - Session Hijacking

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Academic_criterion_8	X
Name	Session Hijacking via XXE injection	X
Exercise	Session Hijacking (see D5.1 for detailed description)	X
Educational objectives	Demonstrating how to use XML External Entities to steal data from a web server	
Underlying indicators	academic_indicator_1, academic_indicator_11, academic_indicator_12	X
Aggregation	The trainee has to perform an XXE injection to steal the session id of another user, the latter will be used to steal a secret text string. The exercise needs to be completed within the time limits.	X
Update frequency	Once per exercise execution	X
Interpretation of the score obtained	If the trainee is able to retrieve both the session id and the secret text string within the time limits, then the exercise is completed correctly	X
Scale		
Uncertainty	No uncertainty applicable to this criterion	X
Storage		
Value, exercise, user, and measurement date		

Table 25. Academic training criterion 8 - Session Hijacking via XXE injection

5.2 Evaluation criteria and indicators for Pilot 2 (transport infrastructure)

Figure 9 provides an overview of the evaluation criteria and their supporting indicators identified for Pilot 2, Transport Infrastructure.

Figure 9 illustrates also the relationship between the evaluation criteria and the indicators. For example, the evaluation criterion TTC1 Event report, SQL injection (described in Table 30) uses the indicators TTI1 Time (described in Table 26), TTI2 Correlation capability (described in Table 27), and TTI3 Forensic capability (described in Table 28). TTC is short for Transport Training Criterion, while TTI is short for Transport Training Indicator.

As illustrated in Figure 9, there are in total 2 evaluation criteria (TTC1 and TTC2), and in total 4 indicators (TTI1 – TTI4) identified for Pilot 2, Transport Infrastructure.

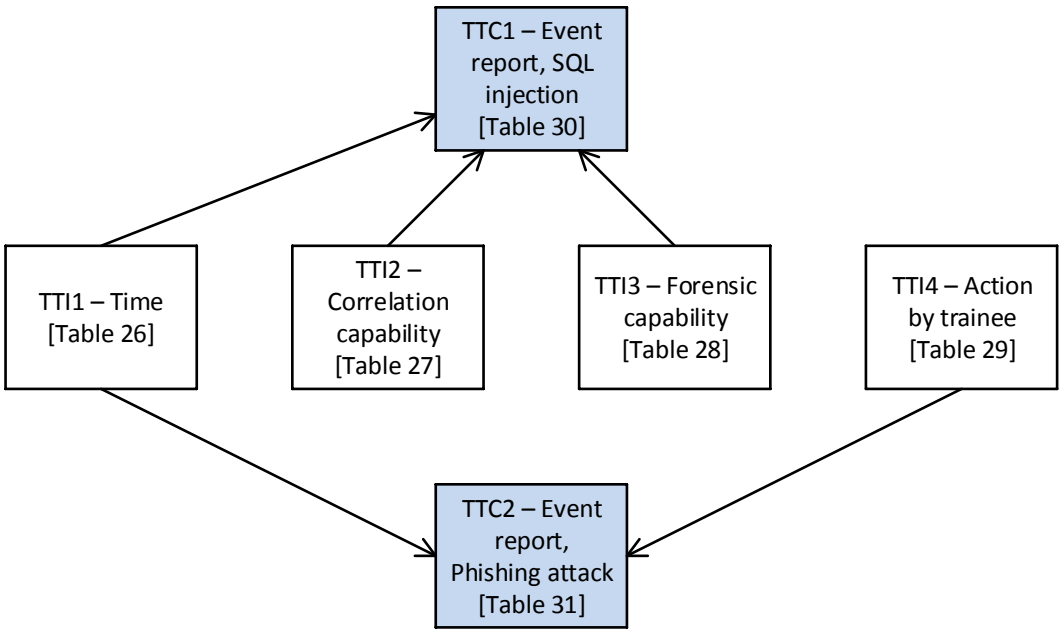

Figure 9. Overview of evaluation criteria and supporting indicators for Pilot 2, Transport Infrastructure

The following tables describe the evaluation criteria (Table 30 and Table 31) as well as the indicators, as illustrated in Figure 9.

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Transport_indicator_1	X
Name	Time	X
Definition	Maximum time to identify the attack	X
Purpose	Measure the time between the attack and its identification	X
Measurement procedure	The trainer will take the starting time when the attack begins and define a known threshold to identify it	
Data source	Trainer's clock's system	
Measurement frequency	Continuously through the exercise until the trainee identifies the attack	X

Attribute	Description of the attribute	Mandatory
Expected change frequency	By the end of the exercise, the time runs out	X
Unit of measure	Seconds	
Interpretation of the value measured	The shortest time the trainee takes to identify the attack, the better performance will be achieved	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 26. Transport training indicator 1 - Time

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Transport_indicator_2	X
Name	Correlation capability	X
Definition	The trainee is able to correlate the events on the network and report anomalous situations	X
Purpose	Check that the trainee is able to report security events	X
Measurement procedure	The trainer will perform an attack and check the response of the trainees	
Data source	Training environment	
Measurement frequency	Continuously through the exercise until the trainee identifies the attack	X
Expected change frequency	After the trainer performs the attack	X
Unit of measure		
Interpretation of the value measured	If the trainee reports the events, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 27. Transport training indicator 2 - Correlation capability

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Transport_indicator_3	X
Name	Forensic capability	X
Definition	The trainee is able to analyze system logs and report anomalous situations	X
Purpose	Check that the trainee is able to detect security events	X
Measurement procedure	The trainer will perform an attack and check the response of the trainees	
Data source	Training environment	
Measurement frequency	One time, after the exercise	X
Expected change frequency		
Unit of measure		
Interpretation of the value measured	If the trainee reports the security breach, then the result is positive	X
Scale		
Uncertainty	Yes, there is the possibility of false positives	X
Storage		
Value, exercise, user, and measurement date		

Table 28. Transport training indicator 3 - Forensic capability

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Transport_indicator_4	X
Name	Action by trainee	X
Definition	The trainee is able to analyze an event and take the right action	X
Purpose	Check that the trainee is able to act correctly	X
Measurement procedure	The trainer will perform an attack and check the response of the trainees	
Data source	Training environment	
Measurement frequency	One time, after the exercise	X

Attribute	Description of the attribute	Mandatory
Expected change frequency		
Unit of measure		
Interpretation of the value measured	If the trainee takes the right action, then the result is positive	X
Scale		
Uncertainty	Yes, there is the possibility of false positives	X
Storage		
Value, exercise, user, and measurement date		

Table 29. Transport training indicator 4 - Action by trainee

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Transport_criterion_1	X
Name	Event report, SQL injection	X
Exercise	SQL Injection (see D5.1 for detailed description)	X
Educational objectives	Learning the importance of reporting incidents	
Underlying indicators	Transport_indicator_1, Transport_indicator_2, Transport _indicator_3	X
Aggregation	Trainees must be able to identify the attack in the shortest time possible and report it	X
Update frequency	Once per attack execution	X
Interpretation of the score obtained	If the trainee was able to quickly identify and report the security event, or is able to detect the attack by analyzing system logs, then the result is positive	X
Scale	Time spent to correctly identify the attack	
Uncertainty	Report of not anomalous activity	X
Storage		
Value, exercise, user, and measurement date		

Table 30. Transport training criterion 1 - Event report, SQL injection

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Transport_criterion_2	X
Name	Event report, Phishing attack	X
Exercise	Phishing attack (see D5.1 for detailed description)	X
Educational objectives	Learning the importance of reporting incidents	
Underlying indicators	Transport_indicator_1, Transport_indicator_4	X
Aggregation	Trainees must be able to identify the attack in the shortest time possible and report it	X
Update frequency	Once per attack execution	X
Interpretation of the score obtained	If the trainee was able to quickly identify and report the security event and take the right actions, then the result is positive	X
Scale	Time spent to correctly identify the attack, action taken	
Uncertainty		
Storage		
Value, exercise, user, and measurement date		

Table 31. Transport training criterion 2 - Event report, Phishing attack

5.3 Evaluation criteria and indicators for Pilot 3 (energy infrastructure)

Figure 10 provides an overview of the evaluation criteria and their supporting indicators identified for Pilot 3, Energy Infrastructure.

Figure 10 illustrates also the relationship between the evaluation criteria and the indicators. For example, the evaluation criterion ETC1 Event report, SQL injection (described in Table 36) uses the indicators ETI1 Time (described in Table 32), ETI2 Correlation capability (described in Table 33), and ETI3 Reputation maintainability (described in Table 34). ETC is short for Energy Training Criterion, while ETI is short for Energy Training Indicator.

As illustrated in Figure 10, there are in total 2 evaluation criteria (ETC1 and ETC2), and in total 5 indicators (ETI1 – ETI5) identified for Pilot 3, Energy Infrastructure.

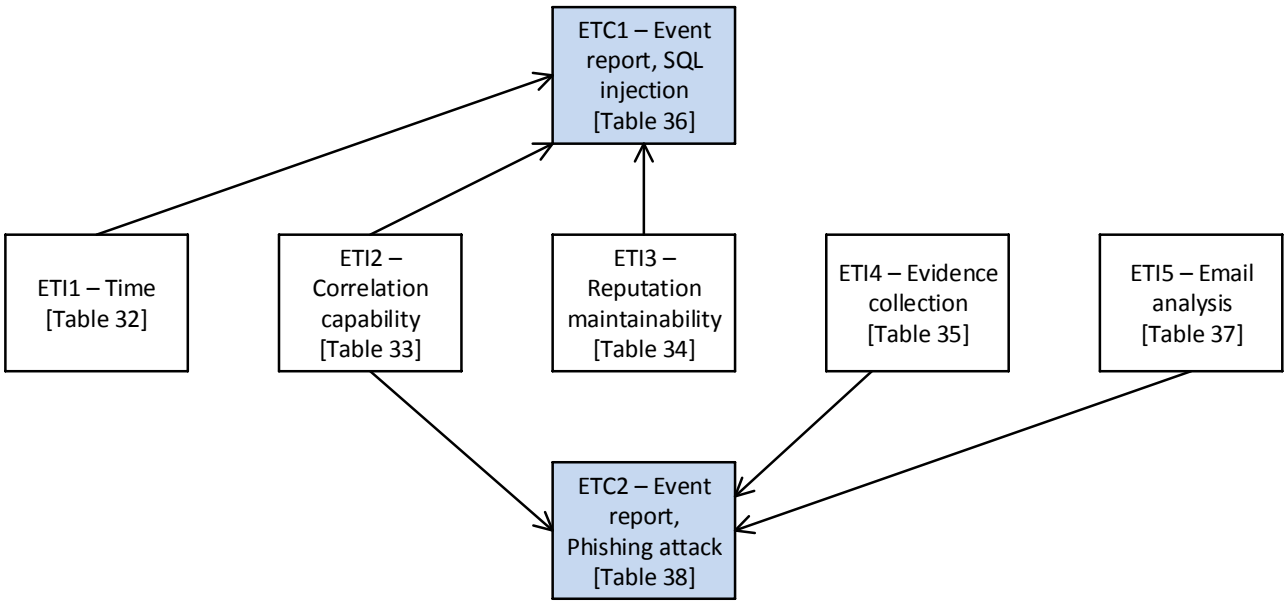

Figure 10. Overview of evaluation criteria and supporting indicators for Pilot 3, Energy Infrastructure

The following tables describe the evaluation criteria (Table 36 and Table 38) as well as the indicators, as illustrated in Figure 10.

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Energy_indicator_1	X
Name	Time	X
Definition	Maximum time to identify the attack	X
Purpose	Measure the time between the attack and its identification	X
Measurement procedure	The trainer will take the starting time when the attack begins and define a known threshold to identify it	
Data source	Trainer's clock's system	
Measurement frequency	Continuously through the exercise until the trainee identifies the attack	X
Expected change frequency	By the end of the exercise, the time runs out	X

Attribute	Description of the attribute	Mandatory
Unit of measure	Seconds	
Interpretation of the value measured	The shortest time the trainee takes to identify the attack, the better performance will be achieved	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 32. Energy training indicator 1 - Time

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Energy_indicator_2	X
Name	Correlation capability	X
Definition	The trainee is able to correlate the events on the network and report anomalous situations	X
Purpose	Check that the trainee is able to report security events	X
Measurement procedure	The trainer will perform an attack and check the response of the trainees	
Data source	Training environment	
Measurement frequency	Continuously through the exercise until the trainee identifies the attack	X
Expected change frequency	After the trainer performs the attack	X
Unit of measure		
Interpretation of the value measured	If the trainee reports the events, then the result is positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 33. Energy training indicator 2 - Correlation capability

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Energy_indicator_3	X
Name	Reputation maintainability	X
Definition	The trainee is able to identify anomalous events on his assets, keeping the reputation as high as possible	X
Purpose	Check that the trainee is able to monitor security events	X
Measurement procedure	The trainer will perform an attack and validate the response on the trainees´ reports	
Data source	Training environment	
Measurement frequency	Continuously through the exercise until the trainee identifies the attack	X
Expected change frequency	After the trainer performs the attack	X
Unit of measure	Percentage	
Interpretation of the value measured	The trainee should avoid decreasing his reputation for 30%, to achieve a positive result	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 34. Energy training indicator 3 - Reputation maintainability

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Energy_indicator_4	X
Name	Evidence collection	X
Definition	The trainee is able to collect evidences during the exercise	X
Purpose	Check that the trainee is able to collect meaningful information about the attack(s) for further investigation	X
Measurement procedure	The trainer will check the information that the trainee was able to collect about the exercise	
Data source	Training environment	
Measurement frequency	Once per exercise	X

Attribute	Description of the attribute	Mandatory
Expected change frequency	Once per exercise, after the trainee finished the exercise	X
Unit of measure		
Interpretation of the value measured	If the trainee was able to collect meaningful information about the attack, then the result will be positive	X
Scale		
Uncertainty	No uncertainty applicable to this indicator	X
Storage		
Value, exercise, user, and measurement date		

Table 35. Energy training indicator 4 - Evidence collection

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Energy_criterion_1	X
Name	Event report, SQL injection	X
Exercise	SQL Injection (see D5.1 for detailed description)	X
Educational objectives	Understand the importance of reporting incidents	
Underlying indicators	Energy_indicator_1, Energy_indicator_2, Energy_indicator_3	X
Aggregation	Trainees must be able to identify the attack in the shortest time possible and report it	X
Update frequency	Once per attack execution	X
Interpretation of the score obtained	If the trainee was able to quickly identify and report the security event, avoiding a decrease of 30% of his reputation, then the result is positive	X
Scale	Time spent to correctly identify the attack	
Uncertainty	Report of not anomalous activity	X
Storage		
Value, exercise, user, and measurement date		

Table 36. Energy training criterion 1 - Event report, SQL injection

Attribute	Description of the attribute	Mandatory
Evaluation indicator ID	Energy_indicator_5	X
Name	Email analysis	X
Definition	The trainee is able to analyse an email and take the right action	X
Purpose	Check that the trainee is able to act correctly upon analysing emails	X
Measurement procedure	The trainer will send emails and check the response of the trainees	
Data source	Training environment	
Measurement frequency	Once per trainee action	X
Expected change frequency	Once per exercise, after the trainee finished the exercise	X
Unit of measure		
Interpretation of the value measured	If the trainee takes the right action, then the result is positive	X
Scale		
Uncertainty	Yes, there is the possibility of false positives	X
Storage		
Value, exercise, user, and measurement date		

Table 37. Energy training indicator 5 - Email analysis

Attribute	Description of the attribute	Mandatory
Evaluation criterion ID	Energy_criterion_2	X
Name	Email report, Phishing attack	X
Exercise	Phishing attack (see D5.1 for detailed description)	X
Educational objectives	Understand the importance of reporting incidents	
Underlying indicators	Energy_indicator_2, Energy_indicator_4, Energy_indicator_5	X
Aggregation	Trainees must be able to identify if each of the emails they receive are phishing attempts or not and report them	X
Update frequency	Once per attack execution	X

Attribute	Description of the attribute	Mandatory
Interpretation of the score obtained	If the trainee was able to correctly identify a phishing email and report it, then the result is positive	X
Scale	Percentage of the number of phishing emails reported over the total number of phishing emails that were sent	
Uncertainty	Report of not malicious emails	X
Storage		
Value, exercise, user, and measurement date		

Table 38. Energy training criterion 2 - Email report, Phishing attack

6. Conclusions

This report describes three main artefacts developed in CYBERWISER.eu as a result of activities in Task T4.3. These artefacts are:

- a method for identifying and specifying evaluation criteria to evaluate the performance of trainees (course participants),
- a set of specific evaluation criteria to be applied in the CYBERWISER.eu courses which will also be used in the pilots defined in WP5,
- a state-of-practice and state-of-art with respect to performance evaluation.

The method for identifying and specifying evaluation criteria is carried out in four steps where the first two steps rely on the identification of learning goals and learning objectives, which is covered as part of course definition process as reported in Deliverable D4.1 [12]. The next two steps are carried out by deriving evaluation criteria and supporting indicators with respect to learning objectives using the templates provided in Section 4.

With respect to evaluation criteria, we have defined:

- in total 8 evaluation criteria and 12 indicators for Pilot 1, academic training,
- in total 2 evaluation criteria and 4 indicators for Pilot 2, transport infrastructure,
- in total 2 evaluation criteria and 5 indicators for Pilot 3, energy infrastructure.

Thus, the evaluation criteria defined have been developed by considering the learning goals and objectives of the exercises for the pilots in WP5 (defined in Task T5.1 and reported in Deliverable D5.1) and the expected corresponding courses to be developed in Task T4.1 for the Intermediate and Advanced offering levels. There are two main reasons to this:

1. The pilots in WP5 are interested in certain learning goals and objectives that require the availability of the technical assets: Performance Evaluator, Economic Risk Evaluator, Countermeasure Simulator, and Vulnerability Assessment Tools.
2. According to the CYBERWISER.eu learning path defined in Deliverable D4.1 [12], courses in which the above technical assets are used will be available in the Intermediate and Advanced offering levels.

The scenarios in which the pilot-exercises will be carried out, will be developed in Task T4.2. The evaluation criteria and indicators defined in this report will be revisited as part of the validation activities in CYBERWISER.eu and adjusted if necessary. Additional evaluation criteria for the courses beyond those that will be used by the pilots in WP5 and that require the abovementioned technical assets will also be developed as part of the validation activities.

In terms of lessons learned:

- The modular approach of the method reported in Section 4 (learning goals → learning objectives → evaluation criteria → indicators) provides a separation of concern and thus facilitates maintainability and scalability.
- The fact that the partners representing the pilots in CYBERWISER.eu managed to individually apply the method in Section 4 to identify evaluation criteria and supporting indicators shows that the method is feasible and works in practice.
- In general, the evaluation criteria and supporting indicators are not generalizable because they must be developed on case basis depending on the learning objectives and supporting training material (courses/exercises). However, in some cases they are generalizable (reusable) such as the Time indicator in Table 6, Table 26, and Table 32.

References

[1] DR. Krathwohl "A revision of Bloom's taxonomy: An overview.", Theory into practice, 2002 Nov 1;41(4):212-8.
[2] Aida Omerovic, Marit Natvig, Isabelle C. R. Tardy "Privacy Scorecard – Refined Design and Results of a Trial on a Mobility as a Service Example". In proceedings of 27th European Safety and Reliability Conference (ESREL' 2017) June 18-22; Portoroz, Slovenia.
[3] Aida Omerovic, Atle Refsdal, Øyvind Rideng "Dynamic Monitoring of Safety Barriers in Petroleum Installations". In proceedings of European Safety and Reliability Association Conference, Amsterdam, ESREL 2013.
[4] Victor R. Basili, Gianluigi Caldiera, H. Dieter Rombach "The Goal Question Metric Approach". Encyclopedia of software engineering, 528-532, 1994.
[5] Terrel Rhodes "Assessing outcomes and improving achievement: Tips and tools for using rubrics". Association of American Colleges and Universities, Washington DC, 2010.
[6] Richard Weiss, Michael E. Locasto, Jens Mache "A Reflective Approach to Assessing Student Performance in Cybersecurity Exercises". In proceedings of the 47th ACM Technical Symposium on Computing Science Education 2016 Feb 17 (pp. 597-602). ACM.
[7] The Kirkpatrick Model. https://www.kirkpatrickpartners.com/Our-Philosophy/The-Kirkpatrick-Model Accessed: 12 August 2019.
[8] Mohamed Faisal Elrawy, Ali Ismail Awad, and Hesham F. A. Hamed: "Intrusion detection systems for IoT-based smart environments: a survey". Journal of Cloud Computing, vol. 7, No. 21. Springer 2018.
[9] Fink, L. Dee. Creating significant learning experiences: An integrated approach to designing college courses. John Wiley & Sons, 2013.
[10] Johns Hopkins, Whiting School of Engineering. Bloom's wheel: https://ep.jhu.edu/files/ep-blooms-wheel.pdf , Accessed: 22 August 2019.
[11] CYBERWISER.eu Project. D2.1 Requirements, Initial Version. November 2018.
[12] CYBERWISER.eu Project. D4.1 Training Material, Initial version. June 2019.
[13] CYBERWISER.eu Project. D2.2 Requirements, Final Version. February 2019.
[14] Association of American Colleges & Universities. VALUE. https://www.aacu.org/value Accessed: 29 August 2019.
[15] Association of American Colleges & Universities. Liberal Education and America's Promise. https://www.aacu.org/leap Accessed: 29 August 2019.
[16] Nmap Security Scanner. Nmap. https://nmap.org/ Accessed: 29 August 2019.
[17] Apache Tomcat. Tomcat 8. https://tomcat.apache.org/download-80.cgi Accessed: 29 August 2019.
[18] AppArmor. AppArmor ("Application Armor"). https://en.wikipedia.org/wiki/AppArmor Accessed: 29 August 2019.
[19] Portswigger Web Security. Burp Suite. https://portswigger.net/burp Accessed: 29 August 2019.