Carnegie Mellon University
Software Engineering Institute

# Self-Assessment in Training and Exercise

Dustin D. Updyke
Thomas G. Podnar
John Yarger
Sean Huff

**October 2024**

**TECHNICAL REPORT**
CMU/SEI-2024-TR-002
DOI: 10.1184/R1/26060911

**CERT® Division**



https://www.sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.



DM24-0611

# Table of Contents

# List of Tables

# Abstract

In this report, we introduce an approach to performance evaluation that focuses on self-assessment. We find that this approach provides both greater information fidelity to satisfy performance assessment objectives and the enhanced realism that cyber operators desired in training and exercise (T&E) activities. We implement a popular incident response tool that enables team members to record their actions and thought processes and facilitate assessing the team's abilities. To further validate our approach, we conducted a survey of participants who used the tool to gather qualitative feedback on its effectiveness. The results of this survey highlight the perceived improvements in realism, the usefulness of self-assessment tools, and the overall impact on team dynamics and individual growth. This combined approach provides valuable insights into team performance, enables best practices to be identified, supports the refinement of mitigation strategies, and fosters actionable feedback for learning. By promoting self-assessment within a realistic T&E environment, this method improves overall team performance in cybersecurity operations by maximizing feedback on individual skills and leadership competencies.

# 1 Introduction

United States Army General Bruce Clarke observed that "An organization does well only those things the boss checks" [Clarke 2021]. This observation highlights the critical role of oversight in organizational performance and the potential for improvement. This principle also underscores the importance of understanding both the essential responsibilities of each role and how to rigorously evaluate them.

From this starting place, our research enhances the alignment of assessments in training and exercise (T&E) events so that actual exercise experiences provide the information capable of assessing the performance of participants, rather than treating performance assessment as an ancillary activity.

This report describes the approach used to integrate self-assessments into team-based cyber warfare exercises. We seek to ensure that exercise methods encourage continuous self-evaluation and oversight, which also helps to achieve consistent managerial vigilance. We integrate assessments into real-world exercises rather than into abstract or gamified elements.

As members of the Cyber Mission Readiness (CMR) team in the CERT® Division at Carnegie Mellon University's Software Engineering Institute (SEI), we take great care to ensure that exercise events are sufficiently realistic facsimiles of what participants will see in the real world. Previously, evaluating individuals and teams during these events had been difficult, forcing participants to separate from the realistic exercise to participate in an assessment.

In our exercise events, we want to remove this separation and incorporate evaluation into the exercise environment. Participants should not need to take additional steps to be assessed successfully. In other words, the assessment would not remove the illusion of realism that we work hard to instill in all aspects of the exercise event.

## 1.1 Assessment Subjectivity

A challenge in cyber assessments is the inherent subjectivity in some evaluations. This subjectivity can be attributed to factors such as the following:

- constraints within the exercise range
- knowledge disparities among participating teams and evaluators
- technical difficulties in comparing participant performance on specific subject matter

For example, gaps in how different teams or assessors understand or apply concepts can lead to discrepancies. These discrepancies are often amplified by the fact that most tasks can be approached in various ways using different tools and techniques. An assessor may have multiple

® CERT is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

methods for solving a technical problem, while the team being evaluated may implement a valid solution that falls outside the assessor's expertise. Additionally, specific technical challenges can make it difficult to measure and compare performance on cybersecurity topics.

These issues may stem from the highly specialized and rapidly evolving nature of cybersecurity, the complexity of the tools and methods used, or the difficulty in creating measurable and objective criteria for cybersecurity work. Ultimately, the challenge is to develop and maintain an assessment method that has the following characteristics:

- is fair and rigorous
- captures the full range of participants' skills and knowledge
- reflects the realities of their work
- enables them to train as they will fight

In our assessments, we strive to evaluate key competencies and qualities, such as the following abilities:

- effectively use relevant cybersecurity tools and techniques
- address and mitigate security challenges efficiently
- work together as a team under pressure and in complex environments
- respond to unexpected situations and evolving threats
- understand the broader context of security measures and their implications

These abilities are critical to our evaluation process for several reasons:

- They ensure robust security postures by verifying that individuals and teams can effectively defend against a wide range of threats.
- They promote continuous improvement. Objective assessments offer actionable feedback that guides future training and skill development, helping participants refine their abilities and address any gaps identified during the evaluation.
- They support real-world applications. Evaluations grounded in practical and transparent criteria ensure that teams can apply their skills effectively in actual security operations, bridging the gap between exercise and real-world execution.

## 1.2 Incident Response

The research we outline in this report led us to review the processes that teams naturally follow throughout their operation that might provide insight into task performance in a more objective and natural way. Incident response is a growing standard practice for both blue and purple cybersecurity teams.[1] Incident response has a maturing set of tools for managing the information

[1] In cybersecurity, a purple team combines the roles of both the red (attackers) and blue (defenders) teams to identify and address security vulnerabilities. The goal is to enhance an organization's overall security by facilitating knowledge sharing and cooperation among the offensive and defensive sides of security.

captured by these teams—from defining an incident to the follow-on activities performed to understand its potential impacts and mitigation. Combining the data that participants captured—including full technical details of the timeline of events executed during the exercise—correlates what happened with how the team responded; this approach tells the story of how the team performed within the exercise.

Over the last decade, we have conducted hundreds of successful high-fidelity, team-based cyber-range exercise events. Our research and first-hand experience led us to use an assessment approach that relies on successful action combined with evaluation of the quality of incident response notes recorded by participants. It is an effective way to understand, supplement, and grow an organization's team-based cyber warfighting capability.

Our team has a history of publishing reports that delineate the value of team-based cyber exercises [Hammerstein 2010], frameworks that provide optimal fidelity and realism (e.g., R-EACTR) [Dobson 2017], the foundations of building a realistic cyber range [Podnar 2021], and supplemental cyber range technical tools (e.g., GHOSTS) [Updyke 2018]. This report follows the spirit of those publications.

We hope the self-assessment we describe in this report provides new approaches for teams contending with the difficulties and shortcomings of traditional cyber assessments.

## 1.3 A Sports Analogy

Throughout the history of our research and development, we have used sports team analogies to highlight how teams, in general, improve their performance over time. Consider the best approach for how sports teams prepare to win on game day. There are several different contexts in play. As individuals, the team's players must be in top physical shape, and they must be able to effortlessly perform the tasks necessary for the position they play. These players must also have positional awareness about where they will play and where their teammates who surround them will play. These individual capabilities are developed through repetitive training; world-class players spend a great deal of time practicing these skills so that each is second nature and requires little thought to execute well.

However, bringing the team together requires the focus to shift from the individual to the dynamics of the team. The players must operate as one cohesive unit, and its members must rely on their intuition to know where a teammate *will be* rather than where they *are*, when to switch the field of play effectively, and so on. This team coordination is often tested in scrimmage games—choosing to play a team like the team it will play when the game counts and treating that scrimmage game as closely to the real thing as possible. A scrimmage game is the best way for the team to know how it will respond on the actual game day.

Carrying this sports analogy to cybersecurity, we refer to *training* as the activities that improve the performance of individual players and *exercise* as the scrimmage game where we assess and improve the team's performance. Just as an infantry unit might exercise in the field, cybersecurity teams exercise on a realistic cyber range.

The field for a cybersecurity scrimmage game is a *cyber range*, which is a fully interactive virtual instance of the information technology (IT) infrastructure for a mid-to-enterprise-level organization that is dedicated to cyber warfare exercise. Cyber ranges include the underlying networks, routers, switches, servers, and workstations we expect to see in real-world implementations. Following our scrimmage analogy, we do everything possible to keep this cyber range as real as possible, from matching the vendors of the toolset our participants use in their daily operations to the toolset's configuration and deployment.

## 1.4 Increasing Realism

Like a good scrimmage game prepares a team for game day, we believe that teams should train in the same way as they intend to fight; teams benefit the most from exercises that most closely mimic the environment and processes they see in their daily operations. As a result, we strive to design our cyber ranges by implementing commercial products that our exercising teams use, and we encourage them to configure these products similarly.

When investigating ways to improve assessments, we hoped to minimize the *out-of-game* experiences that participants are so often forced to use. Our first challenge in the assessment space was determining how to realistically insert some measurement, process, or tool to specifically evaluate how an individual or team is performing. We did this by considering an activity that is happening simultaneously on the cyber range (i.e., potentially a capability that does not exist in an organization's daily operations). Our goal was to find something as unobtrusive as possible and that would reduce the *out-of-game* tasks participants needed to perform.

Often in exercises, there are two universes:

- **In-Game Experiences.** These experiences require the participant to log into the assessment network to perform exercise-related tasks. Activities done within the virtual range environment are typically what we intend to assess.
- **Out-of-Game Experiences.** Also called *game-isms*, these experiences require the participant to step away from realism to perform tasks, answer questions, or complete similar tasks that relate to the exercise or range environment.

The dichotomy of these two exercise universes results in teams having a largely realistic experience until they must step out of that realism to provide information or take a test to support the assessment part of the exercise activity. In an optimal environment, teams would not be required to perform out-of-game tasks or remove themselves from the in-game environment.

# 2 Assessment Considerations

## 2.1 Cyber Range Design

Cyber ranges are often quite complex and large in scope.[2] They often strive to replicate much of what is found in a typical enterprise-class network. Cyber ranges can include enclaves of machines that represent different teams or domains of work (e.g., logistics, medical, operations) and compose multiple machines within each. They can also share common resources (e.g., web and file servers, mail and database systems, directory services). Added to this are an array of routers, proxies, networking tools, and security-specific systems (e.g., sensors, log aggregators).

## 2.2 Informing Assessments

This complexity makes it difficult to be situationally aware of everything a team being assessed might be engaged with at any given point in the exercise. As part of this awareness and tracking challenge, the assessment of participants should have the following characteristics:

- In the best cases, the assessment is *objective* and *quantitative* in nature.
- The assessment provides information beyond what an observer simply sees happening on the cyber range.
- The assessment is *not* based on activity that might not be seen or information that cannot be objectively measured.

In past assessment efforts, we attempted to provide assessors with information about a team that was not observable on the range. This information included what a team was thinking, what it observed but did not act on, and how it responded in ways the assessor could not clearly see. The approaches used to provide this information included quizzes and similar questionnaires or surveys. However, our experience with this approach is that teams would quickly shift their focus toward *scoring* and away from the objectives of the exercise event (i.e., learning or assessing performance).

While these efforts contributed some positives by providing assessors with information that they otherwise would not have had, the negative impacts have always been something we sought to mitigate. Our challenge was how to mitigate the negatives while still providing assessors with information that would better enable them to perform their important duties within the exercise event.

[2] For insight into cyber range construction, functionality, and practical applications, refer to the report *Foundation of Cyber Ranges* [Podnar 2021].

## 2.3 Sources of Assessment Information

Several different roles and responsibilities are necessary for enabling an exercise. Each role covers an important aspect of the assessment:

- **Blue Team Members**: members of the participating team being assessed who likely have different specialties and roles to play within the team
- **White Team Members**: embedded observers or evaluators
- **Black Team Members**: members of the cyber range administration team
- **Red Team Members**: members of the opposing force (If the participating team is defending a defined piece of cyber terrain, there may be an opposing force [OPFOR] [i.e., *red team*] that will attempt to infiltrate that defended territory.)
- **Purple Team Members**: members of the team composed of both red (offensive) and blue (defensive) teams working together to share insights and strategies to improve the cyber range's overall cybersecurity posture (The purple team aims to enhance threat detection and response capabilities through direct cooperation and knowledge exchange.)

All roles provide opportunities to collect information. All roles produce information and artifacts related to the exercise and should be considered in the assessment and the overall output of the exercise. Each role plays an important part in telling the story of what transpired throughout the exercise. However, capturing the intent and thought processes of the participating team members has consistently proved difficult.

## 2.4 Value of Self-Reflection

The actions of the participating teams require attention. It is worth exploring the concept of self-reflection as a potential mechanism for improvement. Can teams learn from their own actions, successes, and mistakes to establish best practices and enhance their performance in future exercises? When properly guided, this concept of self-reflection and self-assessment can be a powerful means of growth and learning. The act of reflecting and documenting such thoughts, actions, and results can also provide invaluable information for the exercise assessors and facilitators as well, which would contribute to a more balanced and in-depth understanding of team performance and identify areas for improvement.

## 2.5 Operational Records

Besides the tools the teams use, the information they routinely capture and use during their operations can provide valuable insights for assessors. This information might include logs from various systems and devices, records of detected incidents, responses to incidents, changes in system configurations, and other similar artifacts. Further, teams often document their standard operating procedures (SOPs), incident response plans, and other policy documents.

Reviewing and comparing the actions taken during an exercise against these established guidelines can provide another layer of objectively assessing the team's performance.

Communication among team members (e.g., emails, chat logs, incident reports) can reveal information about the team's dynamics, collaboration, decision-making process, and overall performance.

## 2.6 Leadership

Finally, a team's leadership and each member's individual proficiency are critical components that can determine the outcome of a battle in any domain, whether on a physical battlefield or in the realm of cyberspace. Just as in kinetic domains, the role of a team leader in a cybersecurity operation is vital. Obalde and Otero describe how important this role is in their readings on combat [Obalde 1998]:

> *In combat, the actions of individual leaders affect the outcome of the entire battle. Squad leaders make decisions and take actions which can affect the operational and strategic levels of war. Well-trained squad leaders play an important role as combat decision makers on the battlefield. Leaders who show initiative, judgment, and courage will achieve decisive results not only at the squad level, but in the broader context of the battle. Without competent squad leaders, capable of carrying out a commander's intent, even the best plans are doomed to failure.*

These insights resonate powerfully within the cybersecurity field. Just as the effectiveness of a military unit hinges on its leaders, the success of a cyber team depends on the competency and skills of its individuals, the strength of its leadership, and the ability of its members to execute the team's strategy and objectives effectively. In cybersecurity, as in combat, well-trained individuals form the bedrock of a team's capabilities, and adept leaders guide the group's actions and strategies, contributing to the wider objectives of the battle. Without this competence, the best-laid plans can fail.

Recognizing these parallels and drawing on the lessons learned from military history, we offer a new perspective on assessing cybersecurity teams and how that assessment can help teams improve—both in terms of leadership and technical proficiency. Our hope is that the approach we detail in this report will prove useful to teams contending with the difficulties and shortcomings of traditional cyber assessment methods.

# 3 Methods

Our developed exercise method empowers participating teams to be the narrators of their own performance stories. By asking them to share not only the actions they undertake but also the reasoning behind those actions, we gain a more comprehensive understanding of their capabilities and strategies. We hope that this method matures beyond the evaluation of team actions to also examine the cognitive process involved, thus providing a richer, more complete picture of individual and team effectiveness.

## 3.1 Incident Response

Central to this approach is an incident response tool. These types of tools are commonly used within security operations centers (SOCs) in many organizations. They facilitate a systematic approach to managing and addressing the security alerts that an organization's intrusion detection systems might raise. Key features of these tools often include ticketing systems for incident management, dashboards for visualization, and features for collaboration and workflow coordination. These functions are critically important within a SOC as, "diagnostic work, i.e. the practice of noticing and categorizing problems, as well as defining the scope of remediation, is a pervasive feature of Information Technology Diagnosis is particularly prevalent during security incident response, one of the primary responsibilities of security practitioners" [Werlinger 2009]. Lastly, "incident response also provides the organization with opportunities to learn" [Ahmad 2019].

By integrating an incident response tool into the assessment process, we repurpose a familiar asset to serve a dual role. Besides being used as a tool for incident response, it also becomes a platform for self-assessment. As participant teams identify threats, formulate responses, and implement mitigations, they document these actions and the associated decision-making processes in the tool.

The incident response tool becomes a live journal for the team as its members document three key aspects of their activities:

- **Detection.** Team members note what they observe or identify as potential threats. This information provides insight into their situational awareness, threat-detection capabilities, and proficiency using security tools. Teams can also receive pre-populated tickets to prompt and direct them during an exercise event. This powerful mechanism keeps teams on task within a time-boxed event.
- **Action**. Team members record the measures they implemented in response to each perceived threat. These measures shed light on their technical proficiency, strategic decision-making abilities, and how they prioritize and manage their responses.
- **Resolution.** Team members document when and how they mitigate a threat or resolve an incident. This information reveals their skills in incident resolution, confidence in their actions, and understanding of the threat landscape.

By leveraging existing incident response tools, we place the narrative power in the hands of the participant teams. As a result, team members are no longer merely subjects of the evaluation but become active participants in the assessment process. This approach opens a new avenue of rich, actionable insights that can effectively gauge their competencies, decision-making abilities, and operational effectiveness.

The strength of our approach lies in combining multiple data sources into a comprehensive narrative. By using team-provided data as a cornerstone, we can contextualize their actions within the broader scope of the exercise. Integrating cyber range information, red team activities, assessor notes, and other pertinent data allows us to craft a robust and detailed account of the day's events. This storytelling approach has several significant advantages:

- **Contextual Understanding.** Understanding the actions of a cybersecurity team in isolation can lead to misleading conclusions. By incorporating additional data points, such as red team activities, we gain a broader perspective on the challenges blue team participants face, the decisions they had to make, and the dynamics of their response.
- **Objective Assessment.** Including information from the cyber range and assessor notes provides an external reference point to the team's self-assessment. This objective data helps mitigate potential biases in self-reporting and ensures a more accurate evaluation.
- **Insightful Correlations.** The amalgamation of diverse data sources allows us to identify correlations and patterns that might otherwise be obscured. For instance, understanding the read team's timeline and actions could provide insight into the team's detection capabilities and response times.
- **Rich Narratives.** The combination of different perspectives—from the team's viewpoint to the red team's actions and the assessor's notes—leads to in a rich and nuanced narrative. This comprehensive account offers a detailed view of the team's performance, the environment in which it operates, and the dynamics of the exercise.

By leveraging this team-provided data and amalgamating it with other information streams, we can deliver a multidimensional, accurate, and insightful account of the cybersecurity exercise, providing a clear picture of a team's performance, strengths, and opportunities for improvement.

## 3.2 Inject Catalog

Our red team activities, commonly referred to as *injects*, are meticulously documented. We include critical aspects, such as the nature of the event, its actions, methods, and underlying objectives. In a U.S. military setting, inject activities are typically aligned with exercising a unit's Mission Essential Tasks (METs), which essentially determine whether a unit can complete its mission. Inject documentation includes expected strategies for mitigating the incident and its potential impacts (e.g., how the team might discover, analyze, and ultimately mitigate the threat or vulnerability).

Having such comprehensive documentation for each inject enables us to more easily identify and evaluate the entries that teams made in the incident response tool. While we initially look for actions that align with expected responses, our process ensures that we do not overlook creative or

unexpected responses. Currently, a team's actions are thoroughly reviewed by hand, allowing us to recognize and evaluate innovative approaches that may deviate from the expected but still effectively mitigate the inject. In fact, these unexpected solutions often lead to valuable insights that improve future inject planning and response strategies.

Our approach to threat profiling has resulted in a catalog of various threats, vulnerabilities, and operational tasks. This catalog serves the following key exercise functions:

- **Benchmarking.** We use our understanding of each threat and its mitigation strategies to establish performance benchmarks for teams. By setting an event timeline for an exercise activity and detailing our expectations for team responses, we can assess and measure team performance against these standards.
- **Identifying Best Practices.** By reviewing entries in the incident response tool, we can pinpoint the effective tactics and strategies teams use to combat specific threats. We can then compile these into a set of best practices for future learning and reference. On the flip side, we can also flag areas for improvement when teams use less successful strategies.
- **Refining Mitigation Strategies.** Team responses give us valuable insights into the practical side of threat mitigation. These insights help us improve our recommended mitigation strategies by making them more effective and applicable to real-world scenarios.
- **Feedback and Learning.** By comparing a team's actions to our expected mitigation strategies, we can provide detailed, actionable feedback. This feedback is crucial to a team's learning process by highlighting the team's strengths and identifying areas for improvement.

By combining detailed threat profiling with team activity data, we can more accurately assess team performance by identifying successful strategies and continuously enhancing our mitigation recommendations.

# 4 Outcomes

Our modular, open sourced assessment application called SEER (System for Event Evaluation Research) provides a centralized dashboard for displaying incident response records from each team. We tested SEER's effectiveness in a series of exercises that covered different scenarios that had a diverse set of associated tasking. The units we tested varied in size, focus areas, and objectives.

SEER is a web application designed to run in an array of range environments, from the Department of Defense's (DoD's) Persistent Cyber Training Environment (PCTE) to the SEI's open source Crucible experimentation and exercise framework. SEER can be deployed onto any network to gather assessment data; it has no dependencies that might affect the ability to run it on a network.

SEER is configured with training objectives from an individual's Training and Readiness (T&R) manual or a team's Mission Essential Task List (METL). An evaluator can view these objectives throughout the exercise and take notes or mark specific performance measures as *go* or *no go*.

Next, to prompt accomplishing these training objectives, we design a realistic scenario-event timeline. A scenario event can be a network or host-based attack (i.e., inject). SEER maintains a catalog with hundreds of exploits that are constantly evolving to reflect current threats. We select and customize applicable injects from this catalog, which contains information about how to execute the inject, which machines it will affect, and the expected responses and mitigations used by the participants to combat it. These injects can be mapped to frameworks (e.g., MITRE Adversarial Tactics, Techniques, and Common Knowledge [ATT&CK] framework).

In the real world, organizations learn about potential compromises through myriad channels. SEER can introduce scenario events via open source threat intelligence distribution platforms such as the Malware Information Sharing Platform (MISP) or other open standards for threat information sharing such as Trusted Automated eXchange of Intelligence Information (TAXII). MISP provides its members with community intelligence about potential threats. Using platforms like these is how our participants can learn about possible new intelligence about threats in their environment (e.g., ransomware, command and control servers) that other organizations have identified.

Once SEER distributes new threat intelligence to MISP, that system automatically updates applications downstream, such as incident response platforms (e.g., TheHive) or threat intelligence management platforms (e.g., ThreatQ). The blue team uses these downstream applications to manage activity during the exercise. For our purposes, we call *these incident response systems*.

During an exercise, the participating teams use the incident response system to track their security operations, including what they notice on the cyber range, explicit tickets assigned to them from higher headquarters, and so on. We might think of the incident response system as a traditional

ticketing system.[3] The team creates some tickets, while other tickets come from upstream systems like the ones we mentioned (e.g., MISP). SEER gathers and organizes the data from all the tickets in the incident response system to provide the evaluator with a centralized view of the unit's operations notes and compares this self-reported performance against the assessment METL.

Each ticket can log information ranging from tracking tags to tasks assigned to specific people (e.g., identifying or mitigating an *indicator of compromise* [IoC]). Tickets can also track other information, including screenshots, images, Internet Protocol (IP) addresses, hashes, and other bits of information the unit captures and tracks. Data about all team-member activity is available with timestamps; activity—what happened and when—can be reviewed throughout the exercise.

Each exercise participant is given a unique login to the incident response system, enabling us to track exercise activity. In terms of assessment, the incident response system is the only tool the exercise participants log into. None of the other systems we discussed require any action from participants.

Within SEER, we map the injects that are part of the exercise to our original training objectives listed in the METL. We define the order in which the injects will play to ensure they realistically integrate with the scenario we constructed.

Exercise administrators and red teams operate using a SEER dashboard throughout the exercise. This dashboard enables them to track information from the incident response system and other systems where information related to a particular inject might be available. Throughout the exercise, they can see all the events that occur within the exercise, their status, and all the information entered by the blue team into the incident response system as events happen.

At the same time, the activity coming in from the incident response system can be validated against the network ground truth in real time by evaluators to substantiate the effectiveness of the blue team's reported actions.

Finally, SEER contains multiple performance reports. As the incident response system and other systems receive activities, that data is overlaid with the event timeline we initially established within SEER. This overlay enables us to correlate when subsequent activities occurred and who was involved. We can even tag certain activities as being *scorable* (e.g., identifying an IoC, identifying a mitigation, meeting a key objective).

We also have a condensed timeline[4] that outlines where a team has spent its time during an exercise, including the time spent per inject. Of course, we can then compare one blue team to another on the same exercises.

---

[3] We are interested in how a participating team might use a security orchestration, automation, and response (SOAR) system in an assessment. SOAR refers to a set of services and tools that automate the prevention and response to cyberattacks by integrating systems, defining response tasks, and developing tailored incident response plans. However, SOAR remains outside the scope of the self-assessment for now.

[4] We also have a more detailed view available.

The initial feedback we have received about the SEER system and teams' use of the incident response system has been positive. Participants found the incident response system to be beneficial to their efforts, and it more closely matched what they strive to do in their day-to-day operations than previous exercise assessment systems. Because of their experience with the exercise, several units reported their continued use of the incident response system and that it improved their daily operations.

## 4.1 Results and Discussion

We compiled results from implementing this new exercise performance assessment approach. Our initial results, shown in Table 1 and Table 2, focus entirely on quantitative metrics that we could derive from receiving a ticket within an incident response system, processing the associated activities,[5] and completing the tasks generated as a result.

*Table 1: Overall Quantitative Results*

| Overall | Delta | Echo |
|---|---|---|
| Participants | 10 | 7 |
| Tickets Closed | 0% | 0% |
| Tasks Closed | 71% | 46% |
| Custom Fields Answered | 39% | 0% |
| Observables Provided | 8 | 10 |
| Attachments Provided | 0 | 17 |
| Ticket Logs | 10 | 6 |
| Entries | 10 | 6 |
| Time Elapsed (in Minutes) | 510 | 851 |

*Table 2: Per-Ticket Quantitative Results*

| Per Ticket | Delta | Echo |
|---|---|---|
| Ticket #1 | | |
| Ticket Closed (in Minutes) | 63 | 235 |
| Tasks Closed | 100% | 100% |
| Custom Fields Answered | 0 | 0 |
| Observables Provided | 1 | 0 |
| Attachments Provided | 0 | 2 |
| Ticket Logs | 13 | 15 |

[5] We include two tickets for illustrative purposes.

| Per Ticket | Delta | Echo |
|---|---|---|
| Ticket #1 | | |
| Entries | 25 | 27 |
| Ticket #2 | | |
| Ticket Closed (in Minutes) | 111 | 191 |
| Tasks Closed | 100% | 43% |
| Custom Fields Answered | 0 | 0 |
| Observables Provided | 5 | 4 |
| Attachments Provided | 6 | 0 |
| Ticket Logs | 11 | 11 |
| Entries | 46 | 28 |

This initial data provides a strictly quantitative comparison between our two exercising teams—Delta and Echo—across various operational metrics without explicitly determining which team performed better. While we might argue that it helps to directly compare each team's efficiency, responsiveness, and thoroughness in handling tickets, that does not offer insight into each team's effectiveness and potential areas for improvement.

We were careful to avoid drawing conclusions based on this data, since it remains difficult to determine outcomes based on factors such as how fast a team operated or how many artifacts they identified. However, we thought we could use this data to identify best practices, pinpoint bottlenecks, and optimize workflows in cyber operations, rather than rank participants or identify that one team performed better than the other.

In subsequent exercises, we had the opportunity to further explore each participant's performance of a set of tasks in a four-hour exercise. For this exploration, we added a weight to each ticket to indicate more intricate or complex tasking.

Using Python, we built a get_score() function that processes the scoring data for each exercise. The function begins by extracting and reading relevant data from the incident response system the team used to log its tickets, tasks, and progress during the exercise. The function then organizes the data by ticket, filters out entries created by range administrators and other non-participants, and removes any duplicates. Next, the function begins overall scoring by completing the following activities:

- Calculate the weight of scores based on a total exercise_score.
- Group the data by task_updated_by and calculate various aggregates (e.g., average points, total score).
- Calculate accuracy and the current score based on weight and scores.
- Sort final data and rank it based on the current score, total score, and total tasks.

Finally, the function converts the Python DataFrame for output, exporting or printing the scoring results by participant from the given exercise to a comma-separated value (CSV) file.

*Table 3: Sample Output of the get_score() Function*

| Participant | Tasks | Task Average | Preliminary | Weight | Exercise Score | Accuracy | Overall Score |
|---|---|---|---|---|---|---|---|
| p_1 | 7 | 5.7 | 40/40 | 32 | 125 | 100 | 12.8 |
| p_2 | 3 | 5 | 15/15 | 12 | 125 | 100 | 1.8 |
| p_3 | 10 | 1 | 10/10 | 8 | 125 | 100 | 0.8 |
| p_4 | 3 | 3.3 | 10/10 | 8 | 125 | 100 | 0.8 |
| p_5 | 2 | 5 | 10/10 | 8 | 125 | 100 | 0.8 |
| p_6 | 2 | 10 | 10/10 | 8 | 125 | 100 | 0.8 |
| p_7 | 1 | 10 | 10/10 | 8 | 125 | 100 | 0.8 |
| p_8 | 1 | 10 | 10/10 | 8 | 125 | 100 | 0.8 |
| p_9 | 1 | 10 | 10/10 | 8 | 125 | 100 | 0.8 |
| p_10 | 1 | 5 | 5/5 | 4 | 125 | 100 | 0.2 |

While the quantitative metrics above provide valuable insights into the operational aspects of our exercise, they do not capture participants' perceptions and experiences, which are crucial for understanding the effectiveness of the exercise and identifying areas for improvement. To address this, we introduced a participant survey designed to gather qualitative feedback and assess perceived improvements over time.

To complement the data derived from the incident response system and further explore participants' perspectives, we developed a Likert-scale survey. We administered this survey to all participants following the exercise and focused on several key areas:

- **Perceived Improvement.** We asked participants to evaluate whether they observed improvements in the realism, relevance, and overall quality of the exercises over time.
- **Assessment Comparison.** We sought feedback on how the assessment approach used in these exercises compared to others they have experienced, particularly regarding the integration of self-assessment and the use of incident response tools.
- **Team Dynamics and Collaboration.** We included questions on how well teams collaborated and whether the exercise enhanced team members' ability to work together under pressure.
- **Effectiveness of Feedback.** We asked participants to assess the usefulness and actionability of the feedback provided during and after the exercise.

The responses from this survey provide a clearer understanding of participants' experiences and offer insights into the effectiveness of our exercise design and assessment methods. In the following tables, we present the survey findings, analyze participant feedback, and discuss the implications for future exercises.

The majority of participants surveyed (Table 4) were relatively new to our exercise program, with 69% having participated for two years or less. Only 13% had been involved for more than five years, indicating a mix of fresh and experienced perspectives in the survey responses.

*Table 4: Participant Experience in Our Cyber Exercise Program*

| How many years have you participated in this exercise program? | Percentage | Count |
|---|---|---|
| First time | 31% | 5 |
| 1-2 years | 38% | 6 |
| 3-5 years | 19% | 3 |
| More than 5 years | 13% | 2 |

Most participants considered themselves at the intermediate (44%) or advanced (38%) levels, with a smaller number identifying themselves at the novice (6%) or expert (13%) levels. This distribution suggests that most of the feedback is from individuals with a solid foundation in cybersecurity.

*Table 5: Participant Cyber Experience Level*

| What would you say your cyber experience level is? | Percentage | Count |
|---|---|---|
| Novice | 6% | 1 |
| Intermediate | 44% | 7 |
| Advanced | 38% | 6 |
| Expert | 13% | 2 |

As shown in Table 6, most participants agreed or strongly agreed that the exercise environment closely mimicked real-world cybersecurity operations, with no participants expressing disagreement. Similarly, 11 participants found the scenarios relevant to their daily responsibilities. The effectiveness of the cyber range setup was also well regarded, with 14 participants agreeing or strongly agreeing that it replicated an enterprise-class network effectively.

The high level of agreement on realism suggests that the exercise design is successfully simulating real-world conditions. However, a small number of participants remained undecided, which may indicate areas where the realism or relevance of specific scenarios could be further enhanced.

*Table 6: Perceptions of Exercise Realism*

| Exercise Realism | Strongly Disagree | Disagree | Undecided | Agree | Strongly Agree |
|---|---|---|---|---|---|
| The exercise environment closely mimics real-world cybersecurity operations. | 0 | 0 | 5 | 6 | 5 |
| The scenarios presented during the exercise are relevant to my daily responsibilities. | 1 | 0 | 4 | 6 | 5 |
| The cyber range setup was effective in replicating an enterprise-class network. | 0 | 0 | 2 | 8 | 6 |

The responses in Table 7 show a generally positive reception to the self-assessment tools, with most participants finding the incident response tool effective for self-assessment and the process itself to be an accurate reflection of their abilities. However, a notable minority (3-4 participants) were undecided or disagreed on the effectiveness and accuracy of the self-assessment process, suggesting that while the tool was generally well received, there were areas that could be improved.

These mixed responses highlight the need for refining the self-assessment tools to ensure they cater to a broader range of participant experiences and expectations. Efforts could focus on making the process more intuitive and ensuring it provides clear, actionable feedback for all participants.

*Table 7: Perceptions of Self-Assessment Integration*

| Self-Assessment Integration | Strongly Disagree | Disagree | Undecided | Agree | Strongly Agree |
|---|---|---|---|---|---|
| The incident response tool used during the exercise allowed me to effectively self-assess my performance. | 1 | 2 | 4 | 4 | 5 |
| The self-assessment process provided actionable feedback that I can apply to future exercises. | 1 | 2 | 3 | 5 | 5 |
| I found the self-assessment process to be an accurate reflection of my abilities and thought processes. | 1 | 1 | 5 | 4 | 5 |
| The integration of assessment within the exercise felt more natural and less disruptive than in other assessments I have participated in. | 0 | 0 | 9 | 2 | 5 |
| Compared to other exercises, the assessment process in this one maintained a higher level of realism and immersion. | 0 | 2 | 5 | 4 | 5 |

The responses in Table 8 reflect strong agreement that the exercise facilitated effective collaboration under pressure, with 15 out of 16 participants agreeing or strongly agreeing. Additionally, 13 participants felt the exercise improved their team's communication and coordination, while 13 also noted that the exercise revealed strengths and weaknesses in team dynamics.

The overwhelmingly positive responses in this category suggest that the exercise is effective in promoting teamwork and highlighting areas for collective improvement. This aspect of the exercise seems to be one of its strongest components.

*Table 8: Evaluation of Team Dynamics and Collaboration*

| Team Dynamics and Collaboration | Strongly Disagree | Disagree | Undecided | Agree | Strongly Agree |
|---|---|---|---|---|---|
| Our team effectively collaborated under pressure during the exercise. | 0 | 0 | 1 | 5 | 10 |
| The exercise helped to improve our team's ability to communicate and coordinate. | 0 | 0 | 3 | 7 | 6 |
| The exercise revealed strengths and weaknesses in our team dynamics. | 0 | 0 | 3 | 6 | 7 |

Responses from Table 9 indicate a general perception of improvement in the quality of exercises over time, though a significant number were undecided. While 11 participants agreed that the exercises became more challenging each year, there is still room for growth in how these improvements are communicated and perceived by participants.

The emphasis on self-assessment as a learning tool was seen as beneficial by most, with 9 participants agreeing or strongly agreeing that it offered a better opportunity for reflection and growth compared to other methods they had experienced.

While there is a positive trend in perceived improvement, the large number of undecided participants suggests that the enhancements may not be consistently recognized. This could be addressed by more clearly demonstrating how each exercise builds on the last and providing explicit comparisons during debriefs.

*Table 9: Perceived Continuous Improvement of Exercises*

| Continuous Improvement | Strongly Disagree | Disagree | Undecided | Agree | Strongly Agree |
|---|---|---|---|---|---|
| I have noticed a consistent improvement in the quality of exercises over the years. | 0 | 0 | 9 | 2 | 5 |
| Each year, the exercises become more challenging and better designed. | 0 | 1 | 6 | 5 | 4 |
| The feedback received from previous exercises has contributed to my growth as a cyber operator. | 1 | 0 | 6 | 3 | 6 |
| The emphasis on self-assessment in this exercise was more beneficial to my learning compared to other assessment methods I've experienced. | 0 | 0 | 7 | 4 | 5 |
| Compared to other exercises, the self-assessment process here provided a better opportunity for reflection and growth. | 0 | 1 | 5 | 6 | 4 |

The majority of participants felt that the exercise effectively identified areas for improvement (12 out of 16), provided a valuable learning experience (11 out of 16), and helped them feel better prepared for real-world cybersecurity incidents (10 out of 16). However, opinions on the overall assessment approach were more mixed. While 10 participants preferred this method over others, some participants remained undecided or disagreed, indicating that there is room for refining the assessment process to better meet participants needs.

The overall effectiveness of the exercise is well supported, though the mixed opinions on the assessment method suggest that future efforts should focus on tailoring the assessment to be more inclusive and effective for all participants. These improvements might involve offering additional training on using the assessment tools or adjusting the tools to be more user friendly.

*Table 10: Overall Effectiveness of the Exercise and Assessment*

| Overall Effectiveness | Strongly Disagree | Disagree | Undecided | Agree | Strongly Agree |
|---|---|---|---|---|---|
| The exercise effectively identified areas where I can improve. | 1 | 1 | 2 | 7 | 5 |
| The exercise provided a valuable learning experience. | 1 | 1 | 2 | 4 | 7 |
| I feel better prepared for real-world cybersecurity incidents as a result of this exercise. | 1 | 2 | 3 | 5 | 5 |
| Overall, I prefer the assessment approach used in this exercise to those I have experienced in other cybersecurity exercises. | 2 | 0 | 4 | 4 | 6 |
| This assessment method is the most effective I have encountered in evaluating my abilities and identifying areas for improvement. | 1 | 2 | 4 | 3 | 6 |

The survey results indicate that the cyber exercises are generally well received, with strengths in realism, team dynamics, and the perceived continuous improvement of the exercises. However, the feedback also highlights areas for refinement, particularly in the integration and effectiveness of self-assessment tools and the overall assessment approach. These insights will guide future iterations of the exercise program while focusing on enhancing participant engagement and ensuring the assessment process is both accurate and actionable for all skill levels.

# 5 Conclusion

In this report, we introduced our findings from our new approach to designing, building, and executing self-assessment within a high-fidelity, team-based cyber exercise event.

These findings show that our exercise and assessment approach enables effective realism and improves overall training value within a wide array of potential cyber warfare exercise scenarios. Survey results indicate that participants generally perceive the exercises to be realistic and relevant, with a majority finding the self-assessment tools to be effective in providing actionable feedback.

Moreover, the feedback gathered suggests that the exercises have positively impacted team dynamics, collaboration, and individual growth, particularly through the emphasis on self-assessment. However, survey results also reveal areas where the assessment process could be further refined to meet the diverse needs of participants, particularly in enhancing the usability and integration of self-assessment tools.

These insights confirm that our framework provides fundamental guidance for exercise developers who aim to create challenging and highly realistic cyber exercise environments. By continuously incorporating participant feedback, we can further refine and enhance the exercise experience, helping team members on their path to becoming elite cyber operators.

We anticipate publishing an ongoing series of reports like this one that will highlight our commitment to realism in exercises, including those that cover the following related topics:

- more technical details about building and operating cyber ranges
- replicating the Internet for training, exercise, and simulation purposes
- properly managing large-scale machine deployments
- mixing cloud and on-premises environments

We also expect to deliver further information about bringing realistic synthetic user activity to life on a cyber range, matching effective red team execution with maximal training value, and conducting elite-level, team-based exercises.

# 6 Next Steps and Future Work

SEER provides a framework for a data-driven assessment of performance. Using our team assessment roadmap, we can automate the detection of successful performance within the environment, integrate data from chats and other collaboration systems, integrate MISP data as a formal part of an exercise, and leverage machine learning algorithms to better automate assessments.

Individual performance assessments involve validating the achievement of learning objectives that are narrower in scope. This narrower scope presents opportunities for more automated assessments that can enable on-demand delivery. We aspire to integrate traditional knowledge checks with automated performance assessments that sense environmental changes on our assessment roadmap.

Future opportunities for this work could include the following:

- Expand the method we outlined in this report to additional teams, contexts, or types of cybersecurity incidents to broaden the applicability of our methods.
- Continue to refine the self-assessment process based on feedback and data from its application. This may involve adjusting the tools used, prompts given to teams, metrics or data gathered, or feedback process in general. Finer precision may provide commanders the ability to compare units and provide a better understanding of what makes a high-performing team.
- Explore how the self-assessment process can be integrated with other systems in organizations outside of exercises (e.g., in human resources or management systems) to better align individual and team development. We assume that team members may feel more empowered and engaged in their development when they directly control some aspects of their performance assessment.
- Based on the self-assessments and assessors' analyses, organizations could investigate creating a catalog of best practices for incident handling and daily operations. This investigation is useful partly because if self-assessments are accurate pictures of current team performance, then perhaps they should align with a team-wide catalog of standard operating procedures.
- Ultimately, each assessment still includes humans as part of the process because of the importance of the output of exercises and assessments. However, to streamline and scale up the process, we should explore opportunities for automation, such as automated data analysis or reporting through further machine-learning techniques.

# References


### [Ahmad 2019]
Ahmad, Atif; Desouza, Kevin C; Maynard, Sean B; Naseer, Humza; & Baskerville, Richard L. How Integration of Cyber Security Management and Incident Response Enables Organizational Learning. *Journal of the Association for Information Science and Technology*. Volume 71. Issue 8. October 2019. Pages 939–953. https://asistdl.onlinelibrary.wiley.com/doi/am-pdf/10.1002/asi.24311

### [Clarke 2021]
Clarke, General Bruce C. *Guidelines for the Leader and the Commander*. Rowman & Littlefield. April 15, 2021. ISBN: 978-0811770200.

### [Dobson 2017]
Dobson, Geoffrey B.; Podnar, Thomas G.; Cerini, Adam D.; & Osterritter, Luke J. *R-EACTR: A Framework for Designing Realistic Cyber Warfare Exercises*. CMU/SEI-2017-TR-004. Carnegie Mellon University, Software Engineering Institute. September 2017. https://insights.sei.cmu.edu/library/r-eactr-a-framework-for-designing-realistic-cyber-warfare-exercises/

### [Hammerstein 2010]
Hammerstein, Josh & May, Christopher. *The CERT Approach to Cybersecurity Workforce Development*. CMU/SEI-2010-TR-045. Carnegie Mellon University, Software Engineering Institute. December 2010. https://insights.sei.cmu.edu/library/the-cert-approach-to-cybersecurity-workforce-development/

### [Obalde 1998]
Obalde Lieutenant M. M. & Otero, Lieutenant A. M. *The Squad Leader Makes the Difference: Readings on Combat at the Squad Level, Volume I*. United States Marine Corps. August 1998.

### [Podnar 2021]
Podnar, Thomas G.; Dobson, Geoffrey B.; Updyke, Dustin D.; & Reed, William. *Foundation of Cyber Ranges*. CMU/SEI-2021-TR-001. Software Engineering Institute, Carnegie Mellon University. May 2021. https://insights.sei.cmu.edu/library/foundation-of-cyber-ranges/

### [Updyke 2018]
Updyke, Dustin D.; Dobson, Geoffrey B.; Podnar, Thomas G.; Osterritter, Luke J.; Earl, Benjamin L.; & Cerini, Adam D. *Ghosts in the Machine: A Framework for Cyber-Warfare Exercise NPC Simulation*. CMU/SEI-2018-TR-005. Software Engineering Institute, Carnegie Mellon University. December 2018. https://insights.sei.cmu.edu/library/ghosts-in-the-machine-a-framework-for-cyber-warfare-exercise-npc-simulation/

**[Werlinger 2009]**
Werlinger, Rodrigo; Muldner, Kasia; Hawkey, Kirstie; & Beznosov, Konstantin. Preparation, Detection, and Analysis: The Diagnostic Work of IT Security Incident Response. *Information Management & Computer Security*. Volume 18. Issue 1. 2010. Pages 26–42. http://lersse-dl.ece.ubc.ca/record/222/files/222.pdf

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. AGENCY USE ONLY | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| (Leave Blank) | October 2024 | Final |

| 4. TITLE AND SUBTITLE | 5. FUNDING NUMBERS |
|---|---|
| Self-Assessment in Training and Exercise | FA8702-15-D-0002 |

6. AUTHOR(S)

Dustin D. Updyke, Thomas G. Podnar, John Yarger, & Sean Huff

| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) | 8. PERFORMING ORGANIZATION REPORT NUMBER |
|---|---|
| Software Engineering Institute<br>Carnegie Mellon University<br>Pittsburgh, PA 15213 | CMU/SEI-2024-TR-002 |

| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSORING/MONITORING AGENCY REPORT NUMBER |
|---|---|
| SEI Administrative Agent<br>AFLCMC/AZS<br>5 Eglin Street<br>Hanscom AFB, MA 01731-2100 | n/a |

11. SUPPLEMENTARY NOTES

| 12A DISTRIBUTION/AVAILABILITY STATEMENT | 12B DISTRIBUTION CODE |
|---|---|
| Unclassified/Unlimited, DTIC, NTIS | |

13. ABSTRACT (MAXIMUM 200 WORDS)

In this report, we introduce an approach to performance evaluation that focuses on self-assessment. We find that this approach provides both greater information fidelity to satisfy performance assessment objectives and the enhanced realism that cyber operators desired in training and exercise (T&E) activities. We implement a popular incident response tool that enables team members to record their actions and thought processes and facilitate assessing the team's abilities. To further validate our approach, we conducted a survey of participants who used the tool to gather qualitative feedback on its effectiveness. The results of this survey highlight the perceived improvements in realism, the usefulness of self-assessment tools, and the overall impact on team dynamics and individual growth. This combined approach provides valuable insights into team performance, enables best practices to be identified, supports the refinement of mitigation strategies, and fosters actionable feedback for learning. By promoting self-assessment within a realistic T&E environment, this method improves overall team performance in cybersecurity operations by maximizing feedback on individual skills and leadership competencies.

| 14. SUBJECT TERMS | 15. NUMBER OF PAGES |
|---|---|
| cybersecurity assessment, self-assessment, training and exercise, T&E activities, realism, cybersecurity operations | 30 |

16. PRICE CODE

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | UL |