Complexity Level of Virtual Simulation Scenarios for Command and Control Behaviors Assessment

Stella Polikarpus*
Estonian Academy of Security Sciences
stella.polikarpus@sisekaitse.ee

Reet Kasepalu
Estonian Academy of Security Sciences
reet.kasepalu@sisekaitse.ee

ABSTRACT

This ongoing evaluates the command-and-control (C2) activities of rescue incident commanders. The study analyzes the assessment results of 382 commanders' C2 behaviors using virtual simulation scenarios in two cases and compares them with other scenarios created through the Collaborative Authoring Process Model for Virtual Simulation Scenarios (CAPM). In the first scenario, involving a chemical leak in a hangar, 60 working incident commanders participated. In the second scenario, 83 commanders responded to a more complex case featuring fire spread and multiple casualties in a care home. Scenario complexity increases with the number of victims, key individuals, and the size of the building. When scenarios become overly complicated for first-level command, the standard ranking of common C2 behaviors is not maintained. To guide the development of future virtual simulation scenarios, recommendations for situational elements should be tailored to each command level, ensuring that the scenario storyline aligns with the appropriate level of likelihood and complexity.

Keywords

Command-and-Control (C2), Effective Command Framework (EC), Collaborative Authoring Process Model for Virtual Simulation Scenarios (CAPM), Virtual Simulation Scenarios (VS)

INTRODUCTION

Assessment tools and measures for command-and-control (C2) agility are identified as a primary research trend (Johansson et al., 2015, p. 3). Effective Command Behavioural Marker Framework (EC) developed by Lamb et al., 2021 focuses on five key behaviours: situational awareness, decision-making, objective setting, action behaviours and review as identified by Launder and Perry, 2014. In the Effective Command tool these five key behaviors are divided into eight assessment sections: perception; comprehension; prediction; decision-making; plan; communication; command; review (Lamb et al., 2021) and this paper we define C2 activities based on those named eight assessment sections. In Estonia, virtual simulation-based training has a history spanning two decades (Polikarpus, Luik, et al., 2023), during which hundreds of virtual simulations have been created for various purposes. Some of these simulations have been used to assess working rescue incident commanders C2 behaviors more than 1,000 times (Polikarpus et al., 2024).

Real-world incidents requiring first-level rescue incident commanders to respond vary widely in both frequency and complexity. These range from routine, relatively simple events—such as trash-bin fires that occur weekly or even daily—to rare, high-consequence incidents like the Grenfell Tower fire in London in 2017, which tragically resulted in 72 fatalities (Gordon, 2018; Dimka, 2023). Incident commanders must be prepared to respond effectively to both ends of this spectrum: frequent, predictable situations and rare, unpredictable, and complex emergencies. Standard Operational Procedures (SOPs) are instrumental in managing routine incidents. However, complex incidents often demand the simultaneous application of multiple SOPs—or, in some cases, an effective response in the absence of any predefined procedures. Interestingly, a recent study found that a simulated "Sinkhole rescue" incident, which required commanders to exercise operational discretion, elicited higher levels of acute stress than a "High-rise fire" scenario that followed established SOPs (Butler et al., 2023). Butler et al. (2023) concluded that "firefighter training

*corresponding author

in SOPs and operational discretion should be augmented alongside personal resilience training, given the impact of stress on health and wellbeing, but also to improve the deployment of SOPs and operational discretion under stress."

While we agree with Butler et al.'s (2023) recommendation, we would like to emphasize a key methodological distinction between their study and the approach adopted in Estonia. In their study, scenarios were presented via moving images displayed in a training room (Butler et al., 2023). In contrast, the Estonian training model utilizes a fully immersive virtual reality environment. During the incident response phase, commanders physically navigate the simulated space, interact with avatars, and communicate via radio, closely mirroring real-life operational conditions (Polikarpus, 2024). We suggest that this level of immersion may significantly influence trainees' cognitive and emotional responses, particularly regarding stress management and decision-making under pressure.

More than a decade ago, scholars already noted that the declining frequency of real-life incidents necessitates the adoption of novel training and assessment methods to mitigate experience gaps among first responders (Lamb et al., 2015). The same author later co-authored a paper emphasizing that live training and virtual simulation-based training should not be directly compared due to their inherently different characteristics (Launder et al., 2015). Nevertheless, subsequent experimental studies sought to examine differences in decision-making across three settings: virtual fires (Experiment 1), simulated fires on a training ground (Experiment 2), and live burns (Experiment 3) (Cohen-Hatton and Honey, 2015). The findings revealed that, regardless of setting, commanders with standard training tended to move directly from information gathering to action. These results support the argument that virtual simulation-based training can be at least as effective as field-based or live training in fostering and assessing C2 behaviors.

Furthermore, the author of the EC has argued that real-world incidents are becoming increasingly complex, highlighting the need for training approaches that prepare personnel for unpredictable and emergent challenges (Lamb et al., 2021). The response of rescue incident commanders to a simulated event can be conceptualized as a temporary system established to preserve life and mitigate incident consequences. These simulated systems can vary in complexity, ranging from simple to complex. Yet, as Northrop (2014) points out, the boundaries between simple, complicated, and complex systems are not always clear-cut and are subject to debate (Northrop, 2014). To our knowledge, this classification has not yet been applied in the context of virtual simulation-based training or the assessment of C2 behaviors among incident commanders.

Complex systems are typically composed of numerous, not necessarily identical, interacting elements whose behavior is governed by rules that may evolve over time (Grabowski and Strzalka, 2008). The interactions between these components tend to be non-linear and may change in undefined ways. A system may be considered complex when it exhibits multiple interacting components, emergent properties, and various levels of embeddedness (Morales-Matamoros et al., 2010). In this paper, we propose using systems classification as an analytical lens to assess the complexity levels of virtual incident simulations (see Table 1). Scenario design should align with the intended command level and the specific C2 behaviors being trained or assessed. As suggested by Reis et al. (2020), training scenarios should, where possible, be based on real-world incidents to ensure authenticity and relevance (Reis and Neves, 2020, p. 37).

In Estonia, a key regulatory requirement mandates that all assessment scenarios for first-level rescue incident commanders be developed using the Collaborative Authoring Process Model for Virtual Simulation Scenarios (CAPM) (Polikarpus et al., 2021). However, CAPM does not stipulate that all scenarios must be based on real incidents. Notably, a practicing rescue incident commander recently stated in a national news outlet that he had not been trained for a particularly complex real-life incident involving the collision of two liquefied natural gas trucks (Vainküla, 2025). This illustrates the potential training gap in preparing for low-frequency, high-impact events. Virtual simulation-based training provides a safe and controlled environment for developing C2 behaviors. In real incidents, the number and nature of interacting elements are inherently unpredictable. In contrast, virtual simulations—especially those authored by trainers—predefine the elements and interactions within a scenario's storyline (Polikarpus, 2024). While this may limit the scenario's alignment with the formal definition of a complex system, from the trainee's perspective, the scenario may still appear highly complex—particularly if key elements are unknown or obscured. Such perceived complexity could, in turn, challenge the validity of C2 behavior assessments, underscoring the need to carefully consider system complexity during scenario design.

All working first- and second-level commanders in Estonia have been regularly assessed using the EC framework since 2016 (Polikarpus and Danilas, 2021). To develop virtual simulations for these assessments, CAPM has been implemented (Polikarpus et al., 2021). The same study compares 22 virtual simulation scenarios and provides recommendations on how scenarios should be authored to effectively measure the situational awareness of rescue incident commanders. The situational elements identified in these virtual simulation scenarios, used to create dynamic decision-making dilemmas, include victims, fire, leaks, and moving traffic (Polikarpus et al., 2021). While we agree that virtual simulations should offer opportunities to train for unexpected situations (Lamb et al., 2021), if

Table 1. Systems Types and Characteristics Based on Morales-Matamoros et al., 2010 and Grabowski and Strzalka, 2008

System Type	*Characteristics*	*Examples*
Simple Systems (where most things are known)	- Few elements - Linear interactions - Predictable behavior	Basic mechanical devices, simple algorithms
Complicated Systems (where we know what we do not know)	- Many elements - Linear but numerous interactions - Predictable with effort	Modern machinery, detailed software programs
Complex Systems (where we do not know what we do not know)	- Many elements - Non-linear interactions - Emergent behavior - High uncertainty	Ecosystems, economies, social networks, weather systems

the assessment results are used to determine whether a commander meets the C2 behavior requirements specified in the occupational qualification standard (Polikarpus and Danilas, 2021), it is essential that the complexity level of virtual simulation scenarios is standardized to ensure fair assessments.

A recent study found that when the same incident commander's C2 behaviors are assessed three times over a period using different virtual simulation scenarios, their three strongest C2 behaviors were consistently Perception, Comprehension, and Plan (Polikarpus et al., 2024). Conversely, their weakest C2 behaviors were Communication, Prediction, and Review. However, earlier studies have demonstrated that the scenario storyline can influence situational awareness assessment results (Polikarpus et al., 2022 and Polikarpus et al., 2021). At the same time, C2 behaviors are considered universal across different incident types and even across various command levels (Launder, 2012; Lamb et al., 2021 and Lamb et al., 2015). A linear relationship is generally assumed: as the complexity level of an incident increases, the required command level must also increase. While simple rescue incidents can and should be managed by first-level commanders together with other partners like police and ambulance, the complicated incidents require higher command levels involvement. Unfortunately, the arrival of higher command levels on accident scene typically takes more time. Consequently, the initial minutes of response, during which first-level commanders apply their C2 behaviors, are critical regardless of the complexity level of the incident.

In real life, the level of incident complexity and location cannot be chosen but this is possible within a virtual environment design. For this study, we selected two virtual simulation scenarios. The first case, KH122 (Scenario 1), involves an incident where two victims need to be rescued from a chemical leak in a hangar building. The second case, KI222 (Scenario 2), features an incident involving the aggressive behavior of a care home resident, multiple casualties, and rapid fire spread. We hypothesize that the scenario storyline where more elements are involved (Scenario 2) influences the assessment results of C2 behaviors.

We posit that a complexity levels exists for real-life events, ranging from simple to complex. Virtual simulation storylines created to train and assess C2 behaviors should reflect this same complexity level in simulated and virtually designed incidents. In the field of trauma patient handling the McGill Simulation Complexity Score (MSCS) is used to objectively rate the difficulty of simulation scenarios (Deban et al., 2023 and Khwaja et al., 2023), however, this system does not suit to rescue incident commanders C2 evaluation. On the one hand, incidents that are highly likely to occur but simple to manage from a C2 perspective are not sufficiently challenging to assess all eight C2 behaviors (Lamb et al., 2021). On the other hand, overly complicated or complex incidents are rare, and extensive safety regulations are in place to prevent such events from occurring. The likelihood of complex incidents remains low. Nevertheless, the ongoing war activity in Europe has highlighted the need for rescue incident commanders to be prepared for far more complex accidents than those typically encountered in peacetime.

Concept of likelihood is closely connected for us to offer authentic assessments for commanders. Authentic assessment should be directly linked to task success, allowing students to demonstrate their knowledge through collaborative, polished performances, with integrated assessment and clear scoring criteria (Herrington et al., 2014). If commander does not believe the storyline and thinks he never needs to respond this kind of accident the assessment is not authentic for him anymore. However, virtual simulation-based training does not inherently ensure believable and engaging roleplay that enhances presence; this largely depends on the skill of those designing the

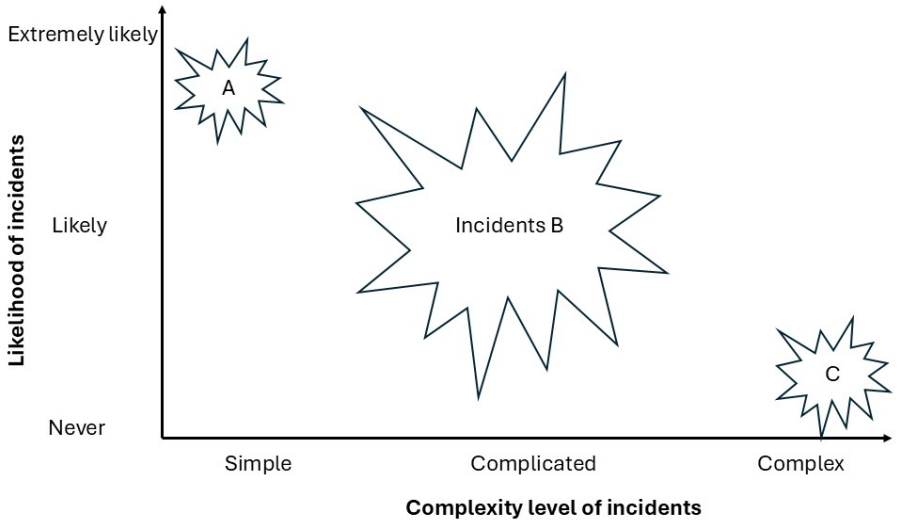

Figure 1. Complexity Level and Likelihood of the Storyline of the Virtual Simulation Scenarios

scenarios (Wijkmark, 2024 p. 148). That is the reason why authors make sure that only these objects are used in scenarios to assess commanders C2 behaviors that are likely to be there in real life as well. There is a need to find balance between complexity and likelihood. Figure 1 illustrates our understanding of how scenario storyline complexity increases while the likelihood of the occurrence of such incidents decreases.

We believe that simple incidents where SOPs can be applied happen daily (see Figure 1 letter A), for example false alarms of automatic fire detection systems. Likelihood of these incidents is very high and theses scenarios seem authentic to commanders, however, implication of SOPs are critical in ensuring effective and efficient rescue operations. SOPs provide a structured approach to managing incidents, ensuring that all responders follow a consistent and coordinated plan. Complex incidents happen very rare (letter C like high-rise building fire) and after incident investigations of these highly complex accidents normally fire safety regulations and SOPs are changed or reviewed. We should aim with virtual simulation-based training for incidents that are likely to happen and they are complicated enough to train and assess first level incident commanders all eight C2 behaviors. However they should not be too complicated or even complex for first level commanders to give them cognitive overload resulting assessment failure, as in real-life they work in teams and together with higher command levels. In Figure 1 letter B is what we should aim for.

Unfortunately, higher task complexity generally leads to increased cognitive load, more errors, and longer response times, which can negatively impact performance ratings. From field of nuclear power the task complexity measures are used to quantify the complexity of tasks and its correlation with human performance to do these tasks. Studies have shown that task complexity measure scores are significantly correlated with performance data such as response times and human error probabilities, indicating that higher task complexity can lead to increased errors and longer response times (Jang and Park, 2022 and Park and Jung, 2007). This seems to be universal principle that more complex situations are prone to greater cognitive load producing more errors in C2 behaviors. However, there is no way it could be tested in real-life because exactly same rescue incidents do not reappear. Nevertheless, if rescue incident is simulated in virtual reality and several commanders can respond to same virtual simulation scenario we can analyze how the storyline of these incidents influences commanders C2 behaviors.

Regrettably, earlier research does not provide guidance on designing virtual simulation scenarios with the right level of complexity to effectively assess all eight C2 behaviors of first-level incident commanders without causing cognitive overload, while ensuring these scenarios are authentic and likely to occur. Our **research aim** is to determine the number and predictability of situational elements and these elements interactions in virtual simulation scenarios 1 and 2, in order to assess the C2 behaviors of first-level incident commanders with a acceptable level of complexity and likelihood.

We propose four research questions:

- RQ1: Which situational elements are present in the compared virtual simulation scenarios that may influence commanders' C2 behaviors?
- RQ2: What are the overall C2 behavior assessment results of compared virtual scenarios?
- RQ3: Which C2 behaviors are most influenced by the virtual simulation scenario storyline?
- RQ4: Which assessment criteria exhibit the most significant differences when the storylines of two scenarios are compared?

The article is organized as follows. First we describe used research and data analyses methods and two selected scenarios. In result section we present findings to each research question. Then we discuss what these findings mean and make suggestions for virtual simulation scenario story lines.

METHODS

We employed a multiple case research design (Hunziker and Blankenagel, 2021, p. 172) as it enables the comparison of rescue incident commanders' C2 behaviors across different incidents presented using virtual reality (Polikarpus, Sarmiento-Márquez, and Ley, 2023). This approach allows us to make informed recommendations regarding the complexity levels of scenarios suitable for training and assessing commanders' C2 behaviors. To answer RQ1 we did document analyses to compare four situational elements (victims, leak, fire, and traffic) (Polikarpus et al., 2021) and to answer other RQs we used C2 behavior assessment results from EC platform.

To select appropriate scenarios, we consulted experts who had utilized all virtual simulation scenarios created with CAPM for first-level commanders between 2022 and 2024 in Estonia. In earlier research where different scenarios created using CAPM were used to assess C2 behaviors (Polikarpus, Sarmiento-Márquez, and Ley, 2023) these two scenarios were not analyzed. The experts identified one scenario they considered relatively simple (Scenario 1, code KH122) and one they deemed more complicated (Scenario 2, code KI222). We then compared the C2 behaviors exhibited by commanders in these two cases with those observed in other scenarios used during the same period, from September 1, 2022, to November 15, 2024.

SAMPLE

We used only the formal assessment results of working first-level rescue incident commanders from the EC database (Polikarpus, 2024). The training process for working rescue incident commanders and the assessment of their C2 behaviors using virtual simulations are described in detail, including curriculum analysis, in Polikarpus et al., 2020, while assessors' training and certification are outlined in Polikarpus and Danilas, 2021. The complete process of scenario creation, validation, and the implementation of virtual simulation-based training for working rescue incident commanders is further elaborated in a doctoral thesis (Polikarpus, 2024).

In Estonia, first-level incident commanders are required to have prior training and work experience as firefighters before completing 30 credit points of vocational education to become a leader of the rescue unit, EstQF Level 5 (Allas et al., 2022). Their C2 behaviors are regularly assessed by certified assessors using virtual simulations developed with CAPM (Polikarpus et al., 2021). These commanders typically work 24-hour shifts at one of the 72 fire stations across the country in Estonia. During incidents, they may collaborate with second-level commanders located in the four regions in Estonia, although it can take more than 60 minutes for a second-level commander to arrive at the incident site.

DATA ANALYSIS

To address RQ1, we applied a document analysis method (Morgan, 2022) of research as it is cost-effective, unobtrusive, presents fewer ethical concerns and offered opportunity to compare the user manuals of the two scenarios. We focused on four situational elements (victims, leak, fire, and traffic) within the storylines of virtual simulation (Polikarpus et al., 2021). For RQ2, we utilized descriptive statistics and constructed a new variable, the behavior score, which represents the average of all eight C2 behaviors assessed. Each criterion was assessed by two certified assessors using a Likert 5-point scale, where a score of 3 represents the threshold level, 1 and 2 indicate performance below the threshold, and 4 and 5 reflect an excellent demonstration of the respective behavior.

To evaluate differences in behavioral scores across scenarios (Scenario 1, Scenario 2, Other Scenarios) RQ3, a one-way ANOVA was chosen as the primary statistical method. The study involved comparing the mean scores of eight C2 behaviors across two independent scenarios and other scenarios mean scores, making ANOVA the

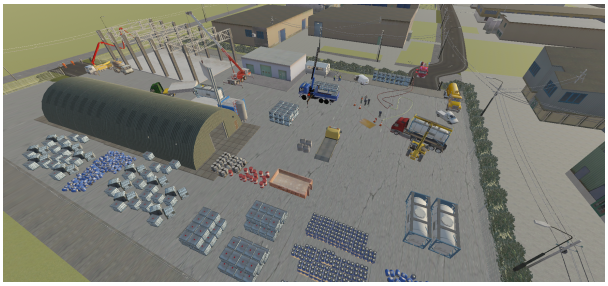

Figure 2. General Situation Picture after Rescuers have Arrived and Set up the Scene in Scenario 1

most suitable method. Before conducting ANOVA, the assumptions of normality and homogeneity of variances were tested to ensure the validity of the statistical analysis. The assumption of normality was assessed using the Shapiro-Wilk test for each scenario and behavior. The test evaluates whether the distribution of scores deviates significantly from a normal distribution. For all behaviors and scenarios, the Shapiro-Wilk test did not indicate significant deviations from normality. Thus, the assumption of normality was satisfied. The assumption of homogeneity of variances was assessed using Levene's test. Levene's test evaluates whether the variances across groups are equal. For all behaviors except Plan, Levene's test indicated that there were no significant differences in variances. For Plan, the test showed a significant difference in variances (p=0.014), indicating a violation of the homogeneity assumption. This meant for the behaviors where the homogeneity assumption was met, a one-way ANOVA was performed to test for significant differences between the Scenario 1, Scenario 2 and Other scenarios. A Welch ANOVA was conducted for the Plan behavior as it is robust to violations of equal variances.

To examine the differences in performance across the Perception behavior (RQ4) which consists of 9 criteria, independent samples t-tests were conducted. This approach was chosen because the analysis focused exclusively on the two scenarios, Scenario 1 and Scenario 2, which exhibited the largest differences in preliminary analyses. Independent samples t-tests were conducted to compare the mean scores between Scenario 1 and Scenario 2 for five variables (Preception behavior assessment criteria) ($q1_6$, $q1_9$, $q1_5$, $q1_7$, and $q1_8$). Due to violations of normality and variance assumptions for $q1_6$, $q1_9$, and $q1_5$, Welch's t-test was employed for these variables, while standard independent samples t-tests were used for $q1_7$ and $q1_8$ as their assumptions of homogeneity of variances were not violated. Effect sizes were calculated to assess the magnitude of the observed differences. For variables analyzed using standard independent samples t-tests ($q1_7$ and $q1_8$), Cohen's d (Cohen, 1998) was used to estimate the effect size, as the pooled variance provided an accurate measure in cases where variances were homogeneous. For variables analyzed with Welch's t-test ($q1_6$, $q1_9$, and $q1_5$), Hedges' g (Hedges, 1981) was employed as it accounts for unequal variances and provides a more robust estimate of effect size in such cases. Both effect sizes were interpreted using conventional thresholds ($d = 0.2$ for small, $d = 0.5$ for medium, and $d = 0.8$ for large effects).

SCENARIO 1: Chemical Leak

We provide details about Scenario 1 based on the scenario's user manual and visual materials created by the authors using XVR OS files.

Description of the Scenario 1 Area

At the edge of a settlement, near the exit area, there is a chemical handling company (see Figure 2). The company's premises are fenced, and the entrance is manned 24/7. Surveillance cameras are installed at the entrance and in key areas within the building, including the site of the incident. The incident involved a chemical leak that occurred during the heating and bottling process. Two employees who attempted to check and mitigate the situation were injured and required help. Beyond life-saving operations, it is critical to identify the chemical substance, determine the location and extent of the leak, and take measures to prevent further escalation of the incident. The rescue operation concludes with transferring the site to the company representative once the chemical substance has been neutralized, the leak has been contained, and the environment is deemed safe for employees to resume operations.

Description of the Scenario 1 Incident

The company handles a chemical (epichlorohydrin), which is fed through a pipeline to the hangar and from there into smaller containers during heating (see the pictures taken by security cameras 3).

During the handling process, an error occurs, resulting in a substance leak from the equipment. As safety regulations do not require personnel to be present at the equipment during bottling, no employees are in the hangar at the time of the incident.

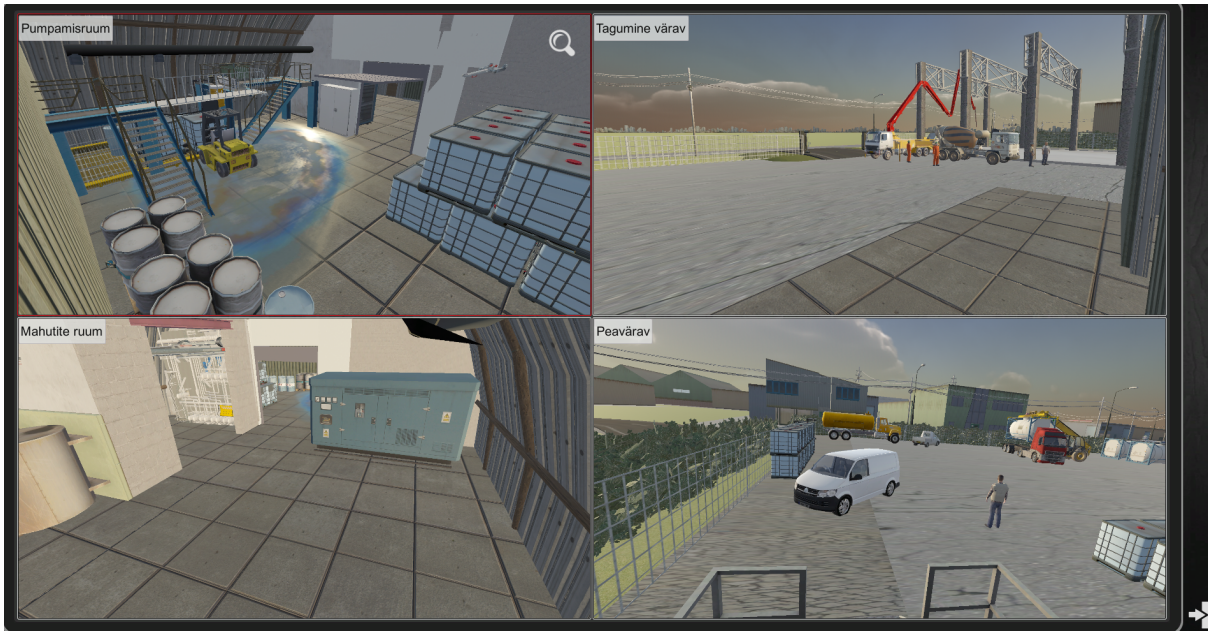

Figure 3. Security Camera Pictures of the First Incident Phase in Scenario 1

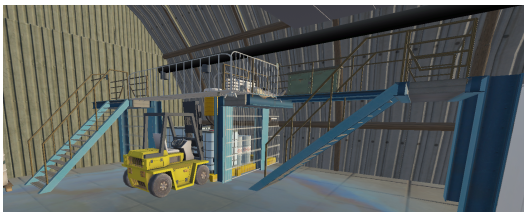

Figure 4. Victim on Loading Ramp: Scenario 1

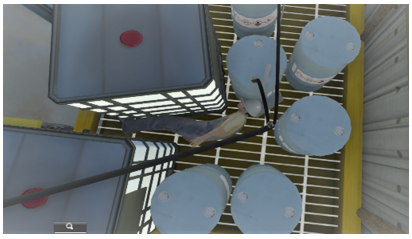

Figure 5. A Victim on the Ground Level Trapped Between Tanks: Scenario 1

The leak is first noticed by a security guard, who observes a cloud of steam near the equipment via surveillance cameras, accompanied by the activation of the automatic fire detection alarm. Although the fire action plan mandates notifying the emergency call center in the event of a fire, the security guard does not observe any visible flames. He silences the alarm and reports the situation to the shift supervisor, who is stationed at the goods delivery point elsewhere on the premises. The shift supervisor, occupied with dispatching goods, instructs two employees to check the situation.

For a reasonable period of time, the shift supervisor attempts to contact the employees by phone to determine what has happened. Neither employee answers. Concerned, the supervisor asks the security guard to display the camera feed of the affected area. The footage reveals that the steam is caused by a leak in the equipment, with a pool of chemical approximately eight square meters in size on the ground, which continues to expand. The employees are not immediately visible in the footage (see Figure 3). The shift supervisor then calls the emergency call center, reporting a chemical leak and the possibility that two employees may have been injured.

The incident commander is expected to organize the rescue of the two casualties, including their decontamination and the administration of first aid as part of life-saving activities. The commander must prioritize which victim requires immediate attention, determine the necessary technical equipment for the rescue, and assess the risk of ignition or further chemical spread. Figure 6 depicts rescue workers administering first aid to a casualty in the decontamination area established by the first team to arrive on the scene.

SCENARIO 2: Care Home Fire

Description of the Scenario 2 Area

It is a special care home consisting of three sections (see Figure 7): the old section, in use since 1993; the new section, built in 2000; and an extension, where construction is ongoing on the second floor while the basement,

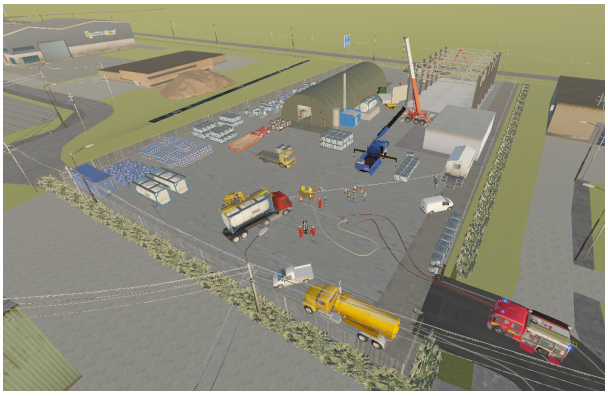

Figure 6. Rescue Team Performing Decontamination of a Casualty in Scenario 1

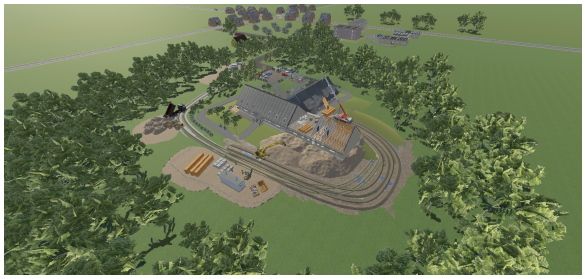

Figure 7. Overview of the Premises in Scenario 2

Figure 8. First Look at the Scene After the Rescue Incident Commander Arrives in Scenario 2

first-floor walls, and ceiling have already been completed. The old and new sections are interconnected, allowing passage between them. The facility accommodates 30 clients with special psychological needs and is staffed by six employees during the day and two at night. Solar panels have been installed on the roof as part of a hybrid system with a network connection. The inverter is located in the basement of the old section, alongside the main electrical panel for both buildings. Additionally, a recently installed Li-ion battery bank (2x70 kWh) is stored on shelves stacked one above the other. These shelves, however, are not secured to the floor or wall.

Description of the Scenario 2 Incident

A special care home is located on the edge of a small settlement and houses approximately 30 residents, with 16 residents accommodated in the old section and 14 in the new section (see Figure 8). An altercation occurs between a resident, Olav, and the staff. During the incident, Olav attacks a caregiver with a knife before fleeing and locking himself in a room on the first floor. The care home staff contacts the police, but they are also unable to access the room. Meanwhile, Olav threatens to set himself and the care home on fire. When the police manage to unlock the door using a key, Olav throws a flammable bottle at them, injuring one officer. While the officers are occupied providing first aid, Olav causes damage in the basement before fleeing in an unknown direction. Fire and smoke begin to spread from the old part of the building to the new part, traveling through the first-floor corridors. Figure 8 illustrates the scene encountered by the commander upon arrival: a police officer and an ambulance team providing first aid outside the building to the injured police officer and a staff member (see also Figure 9).

Figure 9. Other Blue Light Services Already Active on the Scene Upon Rescue Commander's Arrival in Scenario 2

Figure 10. Fire Spreading from the Old Part to the New Part and the Under-Construction Area in Scenario 2

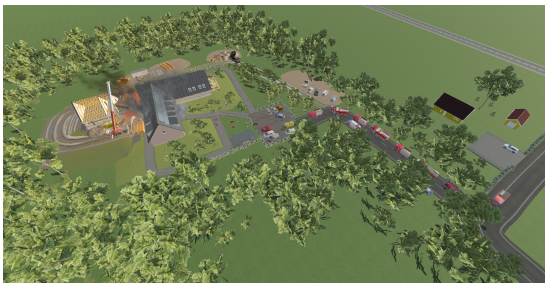

Figure 11. Responding Resources to the Care Home Fire in Scenario 2

Table 2. Comparison of Situational Elements in Scenarios 1 and 2

Scenario	*Victims*	*Fire*	*Leak*	*Traffic*
Scenario 1	2	activated fire alarm, no fire	epichlorohydrin leakage	closed territory, no traffic
Scenario 2	30	rapidly evolving fire	no leak	closed territory, no traffic

RESULTS

Situational Elements in Scenario 1 and 2 (RQ1)

To address our first research question — which situational elements are present in the compared scenarios that may influence commanders' C2 behaviors — we conducted document analysis to deduce differences between "Scenario 1" and "Scenario 2". Four situational elements (see Table 2) we searched for were victims, fire, leak, and moving traffic (Polikarpus et al., 2021).

In Scenario 1, there are two victims endangered by a chemical leak who must be rescued as quickly as possible (see the room where the victims are located, upper-left in Figure 3). One victim is more easily accessible than the other (see Figure 4 and Figure 5). The victim on the loading ramp is unconscious from inhaling chemical fumes but not physically trapped. The second victim, whose clothes are contaminated with the chemical, is lying between chemical tanks, making them harder to locate and rescue (see Figure 5).

In contrast, Scenario 2 involves a more complex storyline, with up to 30 residents, an injured police officer, and an injured staff member (see Figure 9). The scenario is further complicated by the fact that one of the care home residents intentionally set the building on fire after assaulting a police officer and staff. The whereabouts of this individual are unknown to the incident commander, as no key person on-site has this information.

In Scenario 1, the automatic fire alarm is triggered by chemical fumes. The chemical involved, epichlorohydrin, is flammable, but its primary risk is its toxicity. The scenario does not include ignition or fire spread — only the spread of the chemical, which continues until the leak is stopped (e.g., by turning off the pump). The leak is confined to the first floor.

In Scenario 2, fire spreads rapidly due to the building materials (see Figure 10). Therefore, there is clear interaction between three different elements: fire spread possibly producing more victims than initially injured by Olav, Olav unknown location and possibility to attack more people. Construction activities on-site further complicate dynamic risk assessment. Additional hazards include materials from the construction site, solar panels on the roof, and the unpredictably aggressive resident, Olav. These analyses highlight the varying situational elements and complexities that influence the C2 behaviors. To rescue the victims and stop the fire spread in Scenario 2, significantly more smoke-divers are required compared to Scenario 1. Figure 11 illustrates that accommodating all the resources on-site presents a significant challenge due to the narrow entrance to the site.

Table 2 compares the four situational elements of the two scenarios. The table highlights a 15-fold difference in the number of potential casualties between the two sites. The dynamic changes in the scenarios are driven by the chemical leak in Scenario 1 (see Figure 3) and the fire spread in Scenario 2 (see Figure 10). Both scenarios take place in closed sites, eliminating issues from moving traffic (see Figure 6 and Figure 7). Scenario 2 has interlinked elements victims and fire addition to hard to predict human behavior patterns while Scenario 1 does not have linking situational elements, as the number of victims stays the same and leak does not threaten people outside the room.

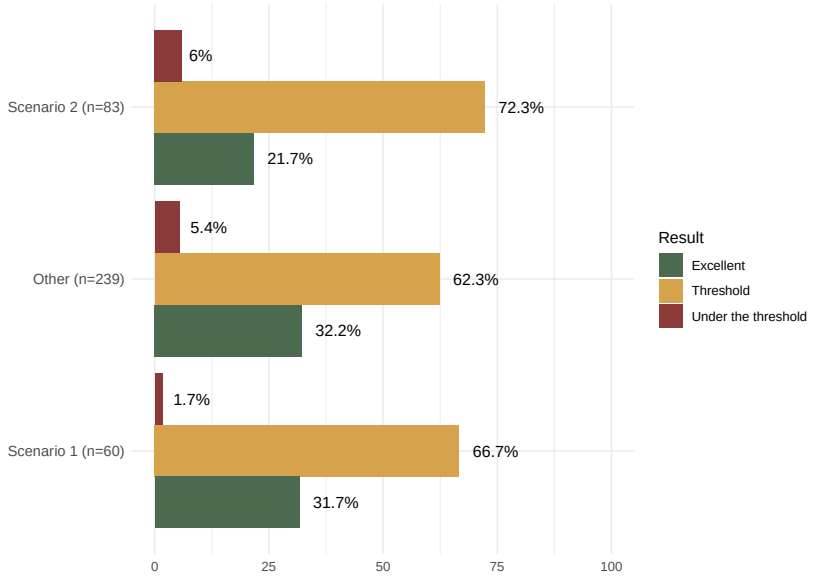

Figure 12. Comparison of Overall Assessment Color (Green-Excellent, Amber-Threshold, or Red-Under the Threshold) Across Scenarios

C2 Behavior Assessment Results (RQ2)

To evaluate the overall C2 behavior assessment results for the compared virtual scenarios, we analyzed the behavior score, which represents the average of all eight C2 behaviors assessed, along with the overall assessment color (green, amber, or red). A green overall assessment indicates that the incident commander demonstrates excellent C2 behaviors. An amber assessment signifies that C2 behaviors meet the occupational qualification standard, representing the threshold level. A red assessment indicates that competencies fall below the threshold.

As illustrated in Figure 12, Scenario 1 is overall easier than both Other scenarios and Scenario 2, as it produced the lowest percentage of results below the threshold (1%). In contrast, 6% of commanders (five individuals) failed Scenario 2, while 72% performed at the threshold level, and only 22% achieved excellent results, highlighting challenges in demonstrating C2 agility.

We compared the overall C2 behavior assessment results across Scenario 1, Scenario 2, and Other scenarios. A one-way analysis of variance (ANOVA) showed that the average behavior scores did not significantly differ between groups, $F(2, 379) = 1.28$, $p = 0.28$. The effect size was very small ($^2 = 0.007$, 90% CI [0.000, 0.021]), indicating that only a negligible proportion of the variance in behavior scores could be attributed to the scenario differences.

These findings suggest that, based on the overall scores, the C2 behaviors demonstrated by first-level commanders were largely consistent across the different scenario types. While we initially hypothesized that scenario-specific complexity or contextual elements (i.e., storyline) might influence assessment outcomes, the current analysis does not provide statistical support for this assumption. It is worth noting that although the descriptive statistics revealed slightly lower average scores in Scenario 2, this trend did not reach statistical significance. Future studies with larger sample sizes or more targeted scenario features may be needed to detect subtle influences of storyline complexity on C2 behavior performance.

C2 Behaviors' Relation to Storylines (RQ3)

To address RQ3 — which C2 behaviors are most influenced by the virtual simulation scenario storyline — we analyzed the mean scores of all eight C2 behaviors (Perception, Comprehension, Prediction, Decision-Making, Plan, Communication, Command, Review) based on nine criteria. Table 3 provides a summary of the comparison of behavior scores across three cases: Scenario 1, Scenario 2, and Others. For each behavior, the mean and standard deviation (SD) are reported. The results indicate that behaviors such as Perception, Comprehension, and Decision-Making show slight variations in scores across scenarios. Scenario 1 generally demonstrates higher mean scores compared to Scenario 2 and Others, suggesting a potential influence of the storyline on these behaviors.

Table 3. Comparison of Behaviors Across Scenario 1, Scenario 2, and Others

Behavior	*Mean (Scenario 1)*	*SD (S1)*	*Mean (Other)*	*SD (O)*	*Mean (Scenario 2)*	*SD (S2)*
Perception	3.38	0.39	3.32	0.34	3.20	0.35
Comprehension	3.32	0.36	3.28	0.36	3.24	0.34
Prediction	3.16	0.32	3.20	0.34	3.12	0.32
Decision-Making	3.16	0.38	3.22	0.36	3.20	0.36
Plan	3.28	0.28	3.25	0.37	3.22	0.36
Communication	3.13	0.38	3.13	0.34	3.12	0.32
Command	3.17	0.35	3.15	0.35	3.07	0.34
Review	3.15	0.31	3.13	0.36	3.08	0.32

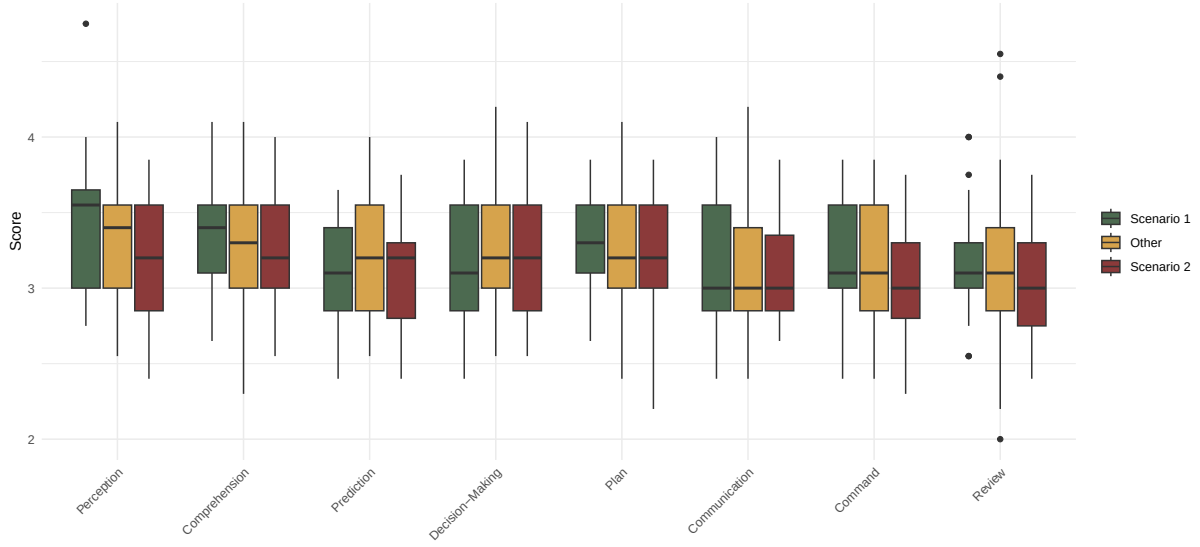

Figure 13. Behavior Scores Across Scenarios: The Central Line Within Each Box Represents the Median, the Box Edges Indicate the Range Where the Middle 50% of the Data Lie, and the Whiskers Extend to Show Most of the Remaining Data. Outliers Are Displayed as Individual Points Beyond the Whiskers.

For instance, the mean score for Perception is highest in Scenario 1 (M = 3.38, SD = 0.39) and lowest in Scenario 2 (M = 3.20, SD = 0.35) Table 3. A one-way ANOVA was conducted to examine whether there were significant differences in scores across the scenarios. The results revealed a significant effect of scenario on Perception scores, $F(2, 379) = 5.25$, $p = 0.006$, with Scenario 1 exhibiting the highest mean score (M = 3.38, SD = 0.39). However, no significant differences were observed for the other behaviors. After identifying significant differences between scenarios for the Perception behavior using ANOVA, a Tukey test was conducted as a post-hoc analysis. This test was used to determine which specific pairs of scenarios exhibited statistically significant differences while controlling for multiple comparisons. The results indicated that the scores for perception in Scenario 2 are significantly lower than in the scores of other scenarios on average, the mean difference (-0.1195) is statistically significant (p = 0.022) . In addition, between Scenario 1 and Scenario 2, the mean difference (-0.1808) is highly significant (p = 0.00754). The results of Scenario 2 are consistently lower compared to both other scenarios and Scenario 2.

To visualize all eight C2 behaviors across all scenarios, we created Figure 13. We used green for Scenario 1, yellow for all other scenarios, and dark red for Scenario 2 to reflect the assumed increasing complexity of the scenario storylines. Our intention with the visual inspection of the medians and interquartile ranges in Figure 13 Behavior Scores Across Scenarios was to gain an initial, scenario-specific overview of how incident commanders' C2 behaviors varied across different types of scenarios. This preliminary exploration served as a basis for identifying potential behavioral patterns that could inform the subsequent, more systematic ranking analysis. From Figure 13, we can observe the median, or midpoint, for each C2 behavior. Notably, the median for Communication behavior is consistent across all scenarios. Beyond that, Other scenarios and Scenario 2 share the same median for Prediction, Decision-Making, and Plan, while Scenario 1 shares its median with Other scenarios for Command and Review behavior. The interquartile range (IQR), representing the range between the 25th and 75th percentiles of assessment results, is the narrowest for Review behavior in Scenario 1, although this behavior also produced three outliers. Three additional outliers were observed for the same behavior in Other scenarios, while the seventh outlier occurred for Perception behavior in Scenario 1.

Table 4. Ranking of C2 Behaviors Across Scenarios (Scenario 1, Others and Scenario 2)

Behavior	*Rank (Scenario 1)*	*Rank (Other)*	*Rank (Scenario 2)*
Perception	1	1	3
Comprehension	2	2	1
Plan	3	3	2
Command	4	6	8
Decision-Making	5	4	4
Prediction	6	5	5
Review	7	7	7
Communication	8	8	6

As the results indicated that the medians for Perception and Comprehension behaviors varied, we further analyzed the ranking of behaviors in Scenario 1, Other scenarios, and Scenario 2 to identify distinct patterns in each scenario. In Scenario 1, Perception was ranked highest, followed by Comprehension and Plan, reflecting a focus on recognizing and understanding the environment and forming plans accordingly (see Table 4). Command was ranked fourth, suggesting a moderate emphasis on directive leadership, while Decision-Making and Prediction occupied intermediate ranks. Review and Communication were ranked lowest, indicating they were less emphasized in this scenario. In Other scenarios, a similar pattern emerged, with Perception, Comprehension, and Plan occupying the top three ranks, although Command dropped to sixth. Review and Communication were again ranked lowest, consistent with Scenario 1. However, Scenario 2 displayed a different pattern, with Comprehension ranked highest and Plan second, reflecting a greater focus on understanding and strategic adaptation in dynamic contexts. Perception fell to third place, while Command and Communication were ranked eighth and sixth, respectively. Despite the differences in ranking order, Review consistently maintained a low ranking across all scenarios, while Decision-Making and Prediction were consistently ranked in intermediate positions, emphasizing their relative importance across contexts. These rankings suggest that while certain behaviors, such as Comprehension and Plan, are emphasized in complicated situations, others, like Perception, are prioritized in more simple or general contexts, reflecting the unique demands of each scenario. Scenario 2 demonstrates a distinct C2 behaviors ranking, with Comprehension taking precedence over Perception and Plan. This shift might indicate a need for deeper understanding and strategic adaptation in more dynamic or complex environments. The low ranking of Review and Communication suggests potential areas for improvement. The consistently low ranking of Review (rank 7 in all cases) may reflect a lack of immediate feedback mechanisms or cognitive overload commander experiences during the dynamic scenario play-through.

Assessment of Perception (RQ4)

To address the question of which assessment criteria show the most significant differences when comparing the storylines of two scenarios, we conducted a detailed analysis of the Perception behavior. This was the only C2 behavior that showed a statistically significant difference between the two compared cases (see Figure 13). Specifically, we examined the nine assessment criteria for Perception behavior as recorded on the EC platform.

The nine criteria (Effective Command, 2025) that were assessed to measure the Perception behavior on certificates were:

- $q1_1$ collection of initial information using relevant incident information
- $q1_2$ initial positioning of the incident commander
- $q1_3$ completion of a 360 (in person or appropriately delegated)
- $q1_4$ identification of appropriate information - critical incident factors
- $q1_5$ gathering information from available sources to gain accurate situation awareness and understanding
- $q1_6$ consideration / utilization of the building layout or risk information relating to incident type
- $q1_7$ procedural info - sourcing of relevant policies/procedures/guidance relating to incident type

- $q1_8$ important information is confirmed
- $q1_9$ communication of the incident situation to other responders

Independent samples t-tests were conducted to compare the mean scores between Scenario 1 and Scenario 2 for five variables ($q1_6$, $q1_9$, $q1_5$, $q1_7$, and $q1_8$). Due to violations of normality and variance assumptions for $q1_6$, $q1_9$, and $q1_5$, Welch's t-test was employed for these variables, while standard independent samples t-tests were used for $q1_7$ and $q1_8$ as their assumptions of homogeneity of variances were not violated. For $q1_6$, Welch's t-test revealed a statistically significant difference between the two scenarios, $t(141) = 2.101$, $p = 0.037$, with Scenario 1 achieving higher scores than Scenario 2. The effect size, calculated as Hedges' g, was $g = 0.42$. Similarly, for $q1_9$, Welch's t-test indicated a significant difference, $t(113.81) = 2.280$, $p = 0.024$, $g = 0.47$. The results for $q1_5$ also showed a significant difference, $t(109.63) = 2.212$, $p = 0.029$, with an effect size$g = 0.45$. For $q1_7$, where assumptions of normality and homogeneity of variances were not significantly violated, the standard independent samples t-test revealed a statistically significant difference, $t(141) = 2.263$, $p = 0.025$. This showed a medium effect size ($d = 0.38$). Finally, $q1_8$ exhibited the strongest effect, $t(141) = 2.643$, $p = 0.009$, with mean scores of 3.42 and 3.16, and an effect size of $d = 0.45$.

The criterion "gathering information from available sources to gain accurate situation awareness and understanding ($q1_5$)" was statistically significantly different for Scenario 1 and 2 because in the first scenario there are two key-persons present to collect all the needed information from (security guard and shift leader), whereas in Scenario 2 there are several staff members present. In Scenario 2, one of the staff members is already injured, the police and ambulance are on site, construction workers, residents and victims outside. Already identifying who has the most up to date information is challenging.

From the comparison between the visualizations of scenarios (RQ1), it was clear that the building plans for Scenario 1 were easier to comprehend than for Scenario 2 and therefore it is logical that criterion "consideration / utilization of the building layout or risk information relating to incident type ($q1_6$)" is statistically significantly different. In addition to this, we highlighted that in Scenario 1 other responders (ambulance, police) arrived after the commander while ambulance would take two victims into hospital. As both scenarios had no issues with moving traffic, police did not have an urgent task in Scenario 1 and was unable to enter the territory anyway because of the toxic substance leak.

Criterion "procedural info - sourcing of relevant policies/procedures/guidance relating to incident type ($q1_7$)" is also very straightforward for Scenario 1 as SOPs exist how to solve chemical accidents and there are no blockages to distract from following them. In Scenario 2 the situation is much more complicated as the alarming message is to assist police in a care home and the threat of setting the house on fire. There is no SOPs available for that situation. Of course there are SOPs for smoke diving, evacuation, dealing with solar panels etc, but the situation demands implementation of several SOPs at the same time and that makes the creation of a safe plan very complicated.

Criterion "important information is confirmed ($q1_8$)" is also much easier for Scenario 1 because the information that needs to be confirmed is the number and location of victims and the substance leak itself while in Scenario 2 it is almost impossible to confirm the number of victims and their exact locations and the smoke and fire spread is changing very fast, thus making that also difficult to confirm.

In addition to this, for Scenario 2 criterion "communication of the incident situation to other responders ($q1_9$)" is very much different, because both are already on the site and have tried to solve the situation and now one police officer is injured and an aggressive person Olav is missing. That addition to bigger building and higher number of victims makes the coordination activities with police and ambulance much harder for commander. Collectively, these findings indicate that the participants in Scenario 1 consistently outperformed Scenario 2 across the assessed criteria, with small to medium effect sizes suggesting meaningful, albeit modest, differences in performance.

DISCUSSION

We aimed to determine the number and predictability of situational elements and these elements interactions in virtual simulation scenarios 1 and 2, in order to assess the C2 behaviors of first-level incident commanders with a acceptable level of complexity and likelihood. However we agree that the notion of a complex system is not delineated yet (Morales-Matamoros et al., 2010). This is the first study on field of rescue incident commanders virtual simulation-based training and assessment addressing this issue. We highlight the need to continue research on how to design using CAPM virtual simulation scenarios for commanders C2 behaviors assessment.

In earlier research four main situational elements of virtual simulation scenarios for rescue incident commanders were identified (see Table 2 Polikarpus et al., 2021). We identified that the situational element "victims" plays a

critical role in determining the complexity level of a scenario. The number of victims is hard to predict both cases, but in Scenario 1 the number is given to commander and it is the same number until the end, while in Scenario 2 the number is changing. Specifically, an increase in the number of potential victims and predicting their physical condition corresponds to a rise in the scenario's complexity level (see Figure 1). This conclusion goes well together with research about trauma patient simulation design where a cardiovascular arrest is planned at any point in the scenario, it results in an automatic maximum score on McGill Simulation Complexity Score (Deban et al., 2023; Khwaja et al., 2023).

Scenario 1 involved a chemical substance leak, while Scenario 2 featured fire spread as the primary situational elements designed to create dynamic challenges for commanders. While the spread of the chemical substance is easily predictable, fire spread is not. In Scenario 1, the chemical leak was confined to a single part and floor of the building, whereas in Scenario 2, the smoke and fire spread across multiple parts and floors. We could identify the interaction of at least two situational elements (victims and fire) in Scenario 2, significantly increasing the scenario's complexity level same time being likely to appear. Therefore, Scenario 2 presented a more complicated situation for the commander to manage compared to Scenario 1. While chemical leaks and fire spread can effectively create significant challenges for first-level commanders, it is important to note that an unusual building layout or multiple floors can further complicate the situation. We also noted that incident caused by human aggressive behavior towards other people in Scenario 2 complicates the simulation even more. In neither scenario was moving traffic used to increase the complexity level; however, the narrow entrance caused by construction works in Scenario 2 made parking the necessary resources more challenging for the commander compared to Scenario 1, where the access route was unobstructed. In Table 2 we provided an overview of the situational elements of both scenarios to answer RQ1.

Overall assessment results (RQ2) also indicated that Scenario 1 was less complicated compared to Scenario 2 (see Figure 12). As the formal assessment results of working commanders play a significant role in their career progression (Polikarpus, 2021), trainers should aim to design scenarios with comparable complexity levels to ensure consistency in assessments. However, if virtual simulation scenarios are designed for training purposes variety of scenarios at different complexity level would be useful to improve the deployment of SOPs and operational discretion under stress (Butler et al., 2023).

When analyzing which C2 behaviors were influenced by the scenarios' storylines (RQ3), we found that the only statistically significant difference between Scenario 1 and Scenario 2 was in the behavior of Perception. At the same time, Scenario 2 demonstrates a distinct ranking of C2 behaviors (see Table 4), where Comprehension takes precedence over Perception and Prediction. In theory, the three levels of situational awareness are defined as Perception (level 1), Comprehension (level 2), and Prediction (level 3) (Endsley, 1995). It has also been previously demonstrated that virtual simulations designed using CAPM for first-level rescue incident commanders allow for the measurement of all named levels (Polikarpus, Sarmiento-Márquez, and Ley, 2023). For a scenario to be suitable for assessing commanders' situational awareness levels it should follow the Endsley defined situation awareness levels and ranking. Other researchers in simulated C2 environment indicate that increasing the volume of information, even when it is accurate and task relevant, is not necessarily beneficial to decision-making performance (Marusich et al., 2016) therefore cognitive overload should be avoided in assessment scenarios.

The C2 behavior results showed only two behaviors with outliers (see Figure 13): Review (six outliers) and Perception (one outlier), none in Scenario 2. This suggests that higher complexity in the scenario storyline does not result in a greater occurrence of exceptional C2 behaviors in assessments. Earlier research has identified Perception, Comprehension, and Plan as the three strongest command behaviors (Polikarpus et al., 2024). Consistently, Scenario 1 and the other scenarios analyzed in this study exhibit the same top three C2 behaviors (see Table 4). Although Scenario 2 also shares these three strongest behaviors, their ranking differs: Comprehension ranks first, followed by Plan and then Perception. In contrast, earlier research found that the weakest command behaviors were Communication, Prediction, and Review (Polikarpus et al., 2024). Scenario 1 aligns with these findings, with the same three weakest behaviors in identical order. However, Scenario 2 differs slightly, with Communication, Review, and Command ranking as the three weakest behaviors, matching the other scenarios from the same time period but in a different order (Table 4). The consistently low ranking of Review and Communication suggests potential areas for improvement, as these behaviors could enhance overall response effectiveness when integrated appropriately. This analysis indicates that the complexity level of the scenario storyline influences the ranking of C2 behaviors. Therefore, to ensure consistency in formal assessments, the complexity level of scenario storylines should be standardized in the future.

When examining the assessment criteria for the C2 behavior Perception (RQ4), we observed findings consistent with RQ1 and earlier research on creating effective assessment scenarios (Polikarpus et al., 2021). Previous research indicated that when a key person is not readily accessible during the life-saving phase of an incident, it becomes

challenging for the commander to establish good situational awareness (Polikarpus et al., 2021). This case study demonstrates that an overabundance of information sources during the initial phase of an incident, as seen in Scenario 2, can lead to cognitive overload for the commander, resulting in a disruption of situational awareness levels. Such complexity in the storyline should be avoided, even though it may occur in real-life scenarios. To mitigate this, it is recommended that a maximum of two key persons be available to the commander upon first arrival at the scene.

A study examining EC results across various countries found that the assessment criterion "$q1_1$ collection of initial information using relevant incident information" is universally regarded as a strength of first-level commanders, while "$q1_6$ consideration / utilization of the building layout or risk information relating to incident type" is identified as a weakness in Perception behavior (Lamb et al., 2021). Based on our findings, we argue that whether these aspects represent strengths or weaknesses for first-level commanders is influenced by the complexity level of the scenario. Additionally, analyses of the scenarios' user manuals (RQ1) and C2 behavior scores (RQ2, RQ3, and RQ4) indicated that responding effectively to Scenario 2 requires the involvement of a higher-level rescue incident commander. This is crucial not only for coordinating the activities of other services on-site but also for ensuring effective cooperation with key persons on site.

In this study, we did not have data on how likely incident commanders perceive these types of incidents (Scenario 1 and 2) to occur in real life (y-axis in Figure 1). However, a previous study indicated that commanders are satisfied with this type of assessment, considering it to be both realistic and authentic (Polikarpus et al., 2020). We recommend that when increasing the complexity level of a scenario, its likelihood should not be compromised. All storylines must remain believable to commanders, ensuring they perceive the incidents as plausible within their own response areas. For this reason, the incident location is always adapted to the commander being assessed (Polikarpus et al., 2020). Based on the analyses of C2 behaviors in Scenario 1 and 2 and their comparison to Other Scenarios, we are confident that both scenarios are class B incidents on the complexity level (see Figure 1).

We would like to emphasize the value of virtual simulation-based training and assessment of C2 behaviors compared to real-life incidents or live training. In real-life incidents, only one first-level commander can arrive first at the scene, and there is typically only one solution to the incident. In contrast, in the virtual environment, multiple commanders can address the same incident, allowing for statistical analysis of assessment results to determine an accepted level of C2 behaviors. However, the development of virtual simulations specifically designed for assessing C2 behaviors requires further study. This paper provides a strong foundation for such future research.

The key takeaways from the study to design virtual simulation scenarios for C2 behaviors assessments are:

- If the scenario has more than one situational element and these elements are interacting to each other it might become already too complicated for commander to gain good level of situation awareness (see Table 4).
- The scenario storyline has a influence to commanders C2 behaviors assessment results and therefore the complexity level of scenarios should be standardized even when CAPM is used in design process.
- Too complicated scenarios, where are several threats including a human antagonistic threat to responders confuses the standard C2 behaviors ranking and might feel more unlikely to commanders.
- For fair virtual simulation-based C2 behaviors assessments scenario likelihood and complexity level score should be developed in future studies.

LIMITATIONS

The main limitation of this study is that all incident commanders originate from the same country and context. However, given that context is considered one of the most critical criteria when studying technology-enhanced training, we argue that conducting multiple case studies within the same socio-economic context is essential (Polikarpus, Luik, et al., 2023). Scenario 1 was solved by fewer commanders (60) compared to Scenario 2 (83). Nonetheless, statistical analyses were conducted to ensure that normality and variance assumptions were met. Errors in performance ratings are often attributed to the cognitive complexity involved in the rating task (Vanhove et al., 2016). Although virtual simulation-based assessment results may be subject to bias, this risk was mitigated by employing two assessors for each evaluation and all assessors are trained and certified. Both assessors reached a consensus on the results. Additionally, all assessors have prior experience as commanders, ensuring the credibility of the evaluations. As the EC framework is utilized in several countries, virtual simulations created using CAPM could be shared in the future to enable comparisons of C2 behaviors across different contexts.

CONCLUSION

The storyline of a virtual simulation scenario has a clear influence on the command and control (C2) behaviors demonstrated by incident commanders. While training scenarios may—and indeed should—vary widely in terms of complexity and likelihood, assessment scenarios require greater consistency. Specifically, they should be designed to reflect comparable levels of likelihood and complexity to ensure fairness and reliability in evaluation outcomes.

As scenario complexity increases—such as through a higher number of casualties or the involvement of multiple agencies—a more advanced command structure may be required earlier in the incident timeline. This contrasts with simpler scenarios, where command structures are typically established more gradually. Consequently, ongoing research into the situational elements of virtual simulation scenarios and their corresponding assessment outcomes is essential. Such research can inform the development of scenario design guidelines that ensure alignment between the scenario storyline and the expected capabilities at each command level. In our analysis, only one of the eight assessed C2 behaviors—Perception—showed a statistically significant difference between the two scenarios. This suggests that the virtual simulation scenarios used to date are, overall, sufficiently challenging to support the effective assessment of first-level commanders' C2 competencies.

REFERENCES

Allas, H., Danilas, K., Laar, R., Sõrmus, K., & Teder, G. (2022). Kutsestandard Päästemeeskonna juht, tase 5.

Butler, P. C., Bowers, A., Smith, A. P., Cohen-Hatton, S. R., & Honey, R. C. (2023). Decision Making Within and Outside Standard Operating Procedures: Paradoxical Use of Operational Discretion in Firefighters. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *65*(7), 1422–1434. https://doi.org/10.1177/00187208211041860

Cohen, J. (1998). *Statistical Power Analysis for the Behavioral Sciences* (2nd).

Cohen-Hatton, S. R., & Honey, R. C. (2015). Goal-oriented training affects decision-making processes in virtual and simulated fire and rescue environments. *Journal of Experimental Psychology: Applied*, *21*(4), 395–406. https://doi.org/10.1037/xap0000061

Deban, M., Iqbal, S., Beckett, A., Posel, N., Fleiszer, D. M., Razek, T., & Khwaja, K. (2023). Virtual trauma patient simulation design using the McGill Simulation Complexity Score (MSCS): a breakthrough in trauma education. *Canadian Journal of Surgery*, *66*(2), E212–E218. https://doi.org/10.1503/cjs.003121

Dimka, N. (2023). Grenfell Tower fire. In *Lessons from grenfell tower* (pp. 16–30). Routledge.

Effective Command. (2025). Level One Effective Command Incident Form.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Gordon, S. G. (2018). Grenfell Tower fire. *The Straight Line*, (6).

Grabowski, F., & Strzalka, D. (2008). Simple, Complicated and Complex Systems - The Brief Introduction. *2008 Conference on Human System Interactions*, 570–573. https://doi.org/10.1109/HSI.2008.4581503

Hedges, L. V. (1981). *Distribution Theory for Glass's Estimator of Effect Size and Related Estimators* (tech. rep. No. 2).

Herrington, J., Reeves, T. C., & Oliver, R. (2014). Authentic Learning Environments. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology: Fourth edition* (pp. 401–412). Springer New York. https://doi.org/10.1007/978-1-4614-3185-5

Hunziker, S., & Blankenagel, M. (2021). *Research Design in Business and Management*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-34357-6

Jang, I., & Park, J. (2022). Determining the complexity level of proceduralized tasks in a digitalized main control room using the TACOM measure. *Nuclear Engineering and Technology*, *54*(11), 4170–4180. https://doi.org/10.1016/j.net.2022.06.018

Johansson, B. J. E., Berggren, P., & Trnka, J. (2015). Research on agility and agile command and control organizations. A review of contemporary literature.

Khwaja, K., Deban, M., Iqbal, S., Alowais, J., Al Bader, B., Deckelbaum, D., & Razek, T. (2023). The McGill Simulation Complexity Score (MSCS): a novel complexity scoring system for simulations in trauma. *Canadian Journal of Surgery*, *66*(2), E206–E211. https://doi.org/10.1503/cjs.002220

Lamb, K., Boosman, M., & Davies, J. (2015). Introspect model: Competency assessment in the virtual world. *ISCRAM 2015 Conference Proceedings - 12th International Conference on Information Systems for Crisis Response and Management*, (August 2005), 235–243.

Lamb, K., Farrow, M., Olymbios, C., Launder, D., & Greatbatch, I. (2021). Systematic incident command training and organisational competence. *International Journal of Emergency Services*, *10*(2), 222–234. https://doi.org/10.1108/IJES-05-2020-0029

Launder, D. (2012). *How do incident managers make decisions in urban fire settings ? An in-depth analysis (dissertation)* (Publication No. September) [Doctoral dissertation, Australian Institute of Business].

Launder, D., Olde, J., Lamb, K., & Link, M. (2015). Simulating Stimulation. *Fire & Rescue, 98*, 3234.

Launder, D., & Perry, C. (2014). A study identifying factors influencing decision making in dynamic emergencies like urban fire and rescue settings. *International Journal of Emergency Services*, *3*(2), 144–161. https://doi.org/10.1108/IJES-06-2013-0016

Marusich, L. R., Bakdash, J. Z., Onal, E., Yu, M. S., Schaffer, J., O'Donovan, J., Höllerer, T., Buchler, N., & Gonzalez, C. (2016). Effects of Information Availability on Command-and-Control Decision Making. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(2), 301–321. https://doi.org/10.1177/0018720815619515

Morales-Matamoros, O., Tejeida-Padilla, R., & Badillo-Piña, I. (2010). Fractal behaviour of complex systems. *Systems Research and Behavioral Science*, *27*(1), 71–86. https://doi.org/10.1002/sres.984

Morgan, H. (2022). Conducting a Qualitative Document Analysis. *The Qualitative Report*, *27*(1), 64–77. https://doi.org/10.46743/2160-3715/2022.5044

Northrop, R. B. (2014, October). *Introduction to Complexity and Complex Systems*. CRC Press. https://doi.org/10.1201/9781439894989

Park, J., & Jung, W. (2007). The appropriateness of TACOM for a task complexity measure for emergency operating procedures of nuclear power plants - A comparison with OPAS scores. *Annals of Nuclear Energy*, *34*(8), 670–678. https://doi.org/10.1016/j.anucene.2007.01.007

Polikarpus, S. (2021). Eesti päästetöö juhid näitavad tehistõelisuses kõrget taset.

Polikarpus, S. (2024). *The Role of Trainers in Designing and Implementing Virtual Simulation-Based Training in Rescue Organisations* [Doctoral dissertation, Tallinn University].

Polikarpus, S., & Danilas, K. (2021). Eesti päästemeeskonna juhtide visiõppepõhise hindamise rakendamine ja tulemused. *Turvalisuskompass*, *1*, 31–54. https://doi.org/https://doi.org/10.15158/4dmc-6f52

Polikarpus, S., Kasepalu, R., & Sarmiento-Márquez, E. M. (2024). From Dynamic Decision-Making Assessments Using Virtual Simulation-Based Training to Targeted Training of Incident Commanders. In W. Seböck, T. J. Lampoltshammer, J. Dugdale, & I. Zeller (Eds.), *Information technology in disaster risk reduction (itdrr2024)*. Springer Nature SwitzerlandAG, Cham (currently in press).

Polikarpus, S., Ley, T., Hazebroek, H., Edgar, G., Sallis, G., Baker, S., & Masip, A. F. (2022). Authoring virtual simulations to measure situation awareness and understanding. In H. Karray, A. D. Nicola, N. Matta, & H. Purohit (Eds.), *Iscram 2022 conference proceedings – 19th international conference on information systems for crisis response and management* (pp. 428–433).

Polikarpus, S., Ley, T., & Poom-Valickis, K. (2020). Developing the Situational Awareness of Incident Commanders: Evaluating a Training Programme using a Virtual Simulation. *Proceedings Estonian Academy of Security Sciences*, *19*, 195–226. https://doi.org/https://doi.org/10.15158/fe4h-ch75

Polikarpus, S., Ley, T., & Poom-Valickis, K. (2021). Collaborative Authoring of Virtual Simulation Scenarios for Assessing Situational Awareness. In A. Adrot, R. Grace, K. Moore, & C. Zobel (Eds.), *Proceedings of the 18th iscram conference* (pp. 229–237). Blacksburg, VA, USA.

Polikarpus, S., Luik, P., Poom-Valickis, K., & Ley, T. (2023). The Role of Trainers in Implementing Virtual Simulation-based Training: Effects on Attitude and TPACK Knowledge. *Vocations and Learning*, *16*(3), 459–486. https://doi.org/10.1007/s12186-023-09322-1

Polikarpus, S., Sarmiento-Márquez, E. M., & Ley, T. (2023). Creation and Use of Virtual Simulations for Measuring Situation Awareness of Incident Commanders. In T. Gjøsæter, J. Radianti, & Y. Murayama (Eds.), *Itdrr2022* (Informatio, pp. 23–38). Springer, Cham. https://doi.org/10.1007/978-3-031-34207-3{_}2

Reis, V., & Neves, C. (2020). Simulations in virtual reality : assessment of firefighters ' decision-making competence. *IE Comunicaciones*, (31), 28–39.

Vainküla, K. (2025). Roolijoodiku jõulusõit: kaks avariid, lennukeeluala ja kõrgeima taseme katastroofioht.

Vanhove, A. J., Gibbons, A. M., & Kedharnath, U. (2016). Rater agreement, accuracy, and experienced cognitive load: Comparison of distributional and traditional assessment approaches to rating performance. *Human Performance*, *29*(5), 378–393. https://doi.org/10.1080/08959285.2016.1192632

Wijkmark, C. H. (2024). *Virtual Reality for Fire and Rescue Service professionals' training* [Doctoral dissertation, Western Norway University of Applied Sciences].

ACKNOWLEDGMENTS

We want to specially thank Tambet Kütt for helping us to conduct this research and sharing with us his expertise. We extend our gratitude to all contributors involved in the creation of the virtual simulation scenarios, the assessors of rescue incident commanders, and the commanders who have diligently worked to resolve the simulated incidents.