TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Zafer Balkan 212289IVCM

BUILDING TASK TEAMS FOR INCIDENT RESPONSE

Master's Thesis

Supervisor: Dr. Ricardo Gregorio Lugo
PhD.

Tallinn 2025

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Zafer Balkan 212289IVCM

VAHEJUHTUMITELE REAGEERIMISEKS TÖÖMEESKONDADE LOOMINE

Magistritöö

Juhendaja: Dr. Ricardo Gregorio Lugo
PhD.

Tallinn 2025

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Zafer Balkan

18.05.2025

Abstract

This study addresses the critical challenge of fostering effective teamwork in *ad hoc* cybersecurity incident response (IR) task teams, which are frequently assembled from diverse personnel with limited shared operational history. Grounded in a comprehensive synthesis of cross-disciplinary literature spanning Team Mental Model (TMM) theory, emergency management, and temporary organizational dynamics, this research investigates the potential of a brief, theoretically-derived TMM intervention to enhance shared understanding. The novelty lies in its application of TMM principles to the unique, high-stakes context of *ad hoc* IR teams and its examination of a minimal-duration (15-minute) intervention.

A careful quasi-experimental design was employed, utilizing two matched *ad hoc* IR teams to minimize bias. Participants, with pre-existing acquaintance but no daily IR collaboration, engaged in pre- and post-intervention tabletop scenarios. The experimental group received a 15-minute TMM-focused intervention (cross-training and strategy briefing), while the control group received standard security training. TMM similarity and accuracy were measured using a multi-faceted approach, including Likert scales, SBERT cohesion analysis of free-text reports, and expert-derived checklists benchmarked against a Gold Standard Model.

While the brief intervention did not yield statistically significant improvements in TMM similarity or accuracy, the study contributes by empirically testing TMM theory in a novel domain and highlighting the complexities of translating theory into short-duration practical interventions for dynamic environments. The rigorous design and multi-method TMM assessment offer methodological insights. Lessons learned inform recommendations for future research, emphasizing the need to explore varied intervention designs, larger samples, and the impact of contextual factors to optimize TMM development in *ad hoc* IR teams.

Keywords: Cybersecurity, Cybersecurity Training, Incident Response, Team effectiveness, Team mental models, Shared mental models

CERCS: T120 Systems engineering, computer technology, S189 Organizational science

Annotatsioon

Vahejuhtumitele reageerimiseks töömeeskondade loomine

Käesolev uuring käsitleb kriitilist väljakutset, kuidas edendada tõhusat meeskonnatööd *ad hoc* küberturvalisuse intsidentidele reageerimise (IR) töörühmades, mis sageli koostatakse erinevatest töötajatest, kellel on piiratud ühine operatiivajalugu. Tuginedes valdkondadevahelise kirjanduse laiaulatuslikule sünteesile, mis hõlmab meeskonna mentaalsete mudelite (TMM) teooriat, hädaolukordade juhtimist ja ajutiste organisatsioonide dünaamikat, uurib see teadustöö lühikese, teoreetiliselt tuletatud TMM-sekkumise potentsiaali jagatud arusaamise parandamiseks. Uudsus seisneb TMM-põhimõtete rakendamises *ad hoc* IR meeskondade unikaalses, kõrge riskitasemega kontekstis ja minimaalse kestusega (15-minutilise) sekkumise uurimises.

Rakendati hoolikat kvaasi-eksperimentaalset ülesehitust, kasutades kahte sarnastatud *ad hoc* IR meeskonda, et minimeerida eelarvamusi. Osalejad, kellel oli varasem tutvus, kuid puudus igapäevane IR-koostöökogemus, osalesid sekkumiseelsetes ja -järgsetes lauaõppustes. Eksperimentaalrühm sai 15-minutilise TMM-keskse sekkumise (ristkoolitus ja strateegiabriifing), samas kui kontrollrühm sai standardse turvakoolituse. TMM-ide sarnasust ja täpsust mõõdeti mitmetahulise lähenemisviisiga, sealhulgas Likerti skaalade, vabatekstiliste raportite SBERT-ühtekuuluvuse analüüsi ja ekspertide koostatud kontrollnimekirjadega, mis olid benchmarkitud kuldstandardi mudeliga.

Kuigi lühike sekkumine ei toonud kaasa statistiliselt olulisi parandusi TMM-ide sarnasuses ega täpsuses, panustab uuring TMM-teooria empiirilisse testimisse uudses valdkonnas ja rõhutab teooria lühiajalistesse praktilistesse sekkumistesse ülekandmise keerukust dünaamilistes keskkondades. Range ülesehitus ja mitmemeetodiline TMM-hindamine pakuvad väärtuslikke metodoloogilisi teadmisi. Saadud õppetunnid annavad soovitusi tulevaseks uurimistööks, rõhutades vajadust uurida erinevaid sekkumisviise, suuremaid valimeid ja kontekstuaalsete tegurite mõju, et optimeerida TMM-ide arengut *ad hoc* IR meeskondades.

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my wife, whose unwavering support, patience, and encouragement have been instrumental throughout this journey. Her belief in me, even during the most challenging moments, gave me the strength to persevere. This work would not have been possible without her steadfast presence by my side.

I would also like to thank Dr. Ricardo Gregorio Lugo for his trust and guidance throughout the development of this thesis. Without his support, this long-running effort might have evolved into a dissertation-length project, rich in scope but lacking the focused direction appropriate for a master's thesis.

List of Abbreviations and Terms

Abbreviation	Meaning
BEC	Business Email Compromise
BERT	Bidirectional encoder representations from transformers
CERT	Computer Emergency Response Team
CIRT	Cyber Incident Response Team
CSIRT	Cybersecurity Incident Response Team
DDoS	Distributed Denial of Service
DFIR	Digital Forensics and Incident Response
ENISA	The European Union Agency for Cybersecurity
IaaS	Infrastructure-as-a-Service
IR	Incident Response
LOTL	Living Off The Land
MaaS	Malware-as-a-Service
MSSP	Managed Security Service Provider
PaaS	Platform-as-a-Service
PhaaS	Phishing as a Service
RaaS	Ransomware-as-a-Service
SaaS	Software-as-a-Service
SBERT	Sentence-BERT
SOC	Security Operations Center
SMM	Shared Mental Model
TMM	Team Mental Model
WARP	Warning, Advice and Reporting Point
XaaS	Everything-as-a-Service

Table of Contents

List of Figures

List of Tables

1. Introduction

Cyber threat environment

According to the European Union Agency for Cybersecurity's (ENISA) 2024 Threat Landscape report[1], cyber threats have become more frequent and complex. Ransomware attacks—where criminals lock or steal data to demand payment—and Distributed Denial-of-Service (DDoS) attacks—which flood systems with traffic to disrupt services—remain among the most common threats, highlighting ongoing challenges to keeping systems running smoothly. Notably, there has been a sharp rise in Business Email Compromise (BEC) incidents. In these attacks, cybercriminals impersonate trusted contacts, such as company executives or partners, to deceive employees into transferring money or sensitive information, often by manipulating existing email conversations or using fake invoices, illustrating heightened exploitation of human factors[2], [3].

Notably, the ENISA report identifies[1] significant advancements in defensive evasion tactics, such as Living Off The Land (LOTL) methods, allowing threat actors to blend seamlessly with normal operational processes. Living off the Land (LOTL) refers to attackers using trusted, built-in system tools—especially LOLBins—to avoid detection by blending in with normal activity. This tactic is effective across various environments (on-prem, cloud, hybrid; Windows, Linux, macOS) and reduces the need for custom malware [4].

Cyber criminals follows the trends of software engineering and IT as well. Software-as-a-service (SaaS) is defined as "a cloud computing service model where the provider offers use of application software to a client and manages all needed physical and software resources" [5]. This allows the users of the service not owning an infrastructure, minimizing the technical capability requirements. Most of the email, instant messaging, CRM and other services are considered SaaS and they are generally paid per operation. This, as-*a-Service* approach extended to other areas like infrastructure as a service (IaaS) or platform as a service (PaaS), depending on the shared responsibilities between the service provider and the user. When it comes to cyber crime, the proliferation of Everything-as-a-Service (XaaS) offerings like Malware-as-a-Service (MaaS), Phishing as a Service (PhaaS) and Ransomware as a Service (RaaS) has been popular and evolving [1]. This lowers the bar to enter into the criminal acts, as cyber crime became pay-per-attack service. Furthermore, AI-based tools, such as FraudGPT, have also emerged, enabling criminals to automate and

enhance phishing and malware deployment efforts [6].

In addition to the attacks around the EU, it is possible to find the impact and global trends in other resources. According to Statista[7], the cost of cybercrime was estimated to be at $8–9 trillion in 2023 and projected to exceed $15 trillion by 2029. In the USA, the FBI received 880,418 cybercrime complaints in 2023, with losses surpassing $12.5 billion [2]. The financial and healthcare sectors face especially high risks. Healthcare breach costs averaged $10.93M in 2023 [8]. On the other hand, the geopolitical aspects had a meaningful impact on cyberspace: part of the Russian invasion of Ukraine, over 4,300 cyber incidents targeted Ukrainian infrastructure in 2024 [9].

What is Incident Response?

An incident in IT service management is defined as "an unplanned interruption to a service, or reduction in the quality of a service" [10]. However, this definition generally focuses on availability, though vaguely it may include confidentiality and integrity aspects, which are to be ensured for information security [11]. Therefore incident needs to be defined in information security or cybersecurity perspective. An incident is defined as [12]:

> An occurrence that actually or potentially jeopardizes the confidentiality, integrity, or availability of an information system or the information the system processes, stores, or transmits or that constitutes a violation or imminent threat of violation of security policies, security procedures, or acceptable use policies.

Therefore, responding to the incidents carry a specific meaning. Aligning with the definition above, incident response (IR) in cybersecurity refers to "[t]he remediation or mitigation of violations of security policies and recommended practices"[13]. It encompasses the technologies and procedures for managing cyber threats and breaches in order to limit damage and restore normal operations. The National Institute of Standards and Technology (NIST) emphasizes that a formal IR capability is crucial for "rapidly detecting incidents, minimizing loss and destruction, mitigating the weaknesses that were exploited, and restoring IT services" [13].

The increasing sophistication of cyber threats, as mentioned in Section 1 Cyber Threat Environment, combined with the complexity of IR activities, highlights the critical need for preparedness as a team. However, working as a team in the response may be challenging. According to Kleij *et al.* [14]:

> From the general teams literature, it is known that teams are not easily implemented, that the creation of a team of skilled members does not ensure success, and that teamwork does not just happen. In fact, many teams never reach their full potential, and many fail altogether.

And therefore, the teams do not exist in a vacuum; they must be built. Throughout the years, it has been found out that effective IR training updates technical knowledge and promotes consistent methodologies [13], enhances readiness and decision-making through realistic simulations [15], and prepares management across legal, communications, and operational roles through cross-functional training [16]. There are common strategies available in literature to be explained within this research.

Task teams

Through time IR as a concept evolved in time. According to NIST SP 800-61r3, IR was once the exclusive domain of specialized internal teams, contemporary practices reflect a broader understanding of organizational interdependence [13]. Incident handlers remain vital, but the success of IR now hinges on the coordinated involvement of a wide range of internal and external actors. These participants span diverse functions and may be geographically distributed, with roles and responsibilities that vary not only between organizations but also across different types of incidents.

Such configurations align with the concept of a temporary organization, defined as "a set of organizational actors working together on a complex task over a limited period of time" [17]. They may also be described as "*ad hoc teams* which are gathered to solve a specific problem" [18], or as task-focused teams formed for a limited duration [19]. While the dynamics of these temporary, *ad hoc*, task-specific teams have been examined in domains such as healthcare, emergency management, and aviation, their role in cybersecurity IR remains underexplored. This study focuses specifically on the team-building dimension of such temporary structures, although related aspects - including leadership, management styles, communication practices, and coordination mechanisms - have been addressed in the broader literature.

Building task-teams for IR

This study seeks to address team building in diverse, cross-functional, and temporary units by investigating how targeted training interventions can improve the development of Team Mental Models (TMM). Specifically, the research examines whether improving TMM

similarity and accuracy leads to more effective team performance during IR scenarios. By focusing on the cognitive and shared understanding aspects of team function rather than solely on procedural or structural mechanisms. This study aims contribute to the broader literature on team cognition and provides empirically grounded insights into enhancing the performance of IR teams operating in complex, dynamic, and time-sensitive environments.

2. Research Questions

2.1 The Main Research Question (MRQ) and Sub-Research Questions (SRQs)

MRQ – To what extent does targeted training on Team Mental Models (TMM) improve the effectiveness of *ad hoc* task teams in cybersecurity incident response?

The question is broken down into two sub-research questions (SRQs):

SRQ1 – How does TMM-specific training affect the similarity of mental models among incident response team members compared to traditional security training?

SRQ2 – To what degree does TMM-specific training improve the accuracy of team mental models in incident response scenarios as evaluated by domain experts?

2.2 Rationale for Research Questions

2.2.1 Main Research Question

The MRQ addresses the core hypothesis that targeted TMM training can improve incident response effectiveness. It is specific to the experimental intervention, measurable through the collected data, focused on the practical application, and aligned with the experimental design.

2.2.2 SRQ1: Similarity Dimension

This question focuses on the first key dimension of TMM - similarity or sharedness across team members. It allows for direct comparison between experimental and control groups, can be measured through both Likert scale responses and semantic similarity analysis, and addresses a fundamental aspect of TMM theory.

2.2.3 SRQ2: Accuracy Dimension

This question addresses the second key dimension of TMM - accuracy of the shared understanding. It recognizes that similarity alone is insufficient, leverages the expert evaluation component of the methodology, and provides a quality assessment beyond mere convergence

These research questions form a comprehensive framework that addresses both the theoretical aspects of TMM (similarity and accuracy) and their practical implications for incident response effectiveness.

3. Literature Review

This chapter reviews the academic literature relevant to building effective *ad hoc* task teams for cybersecurity incident response (IR), with particular attention to the role of Team Mental Models (TMMs), also referred to as Shared Mental Models (SMMs). The review begins by outlining the defining characteristics of *ad hoc* IR teams, drawing comparisons with emergency response teams from other domains to establish a working definition suitable for the cybersecurity context. It then turns to the theoretical foundations of TMMs, examining their core dimensions—including task, team, temporal, and complexity-related aspects—as well as their underlying mechanisms, their relationship to team processes, and empirical findings linking TMMs to team effectiveness. Special emphasis is placed on the unique challenges and requirements associated with TMM development in the context of *ad hoc* IR teams. The chapter concludes by discussing evidence-based strategies for fostering TMMs and addressing key methodological and conceptual issues in TMM research, including a synthesis of commonly used measurement approaches.

3.1 Defining and Characterizing Incident Response Teams

As mentioned in the Chapter 1 Introduction, while cybersecurity incident response is often associated with designated security teams (e.g., SOCs, CERTs, or CSIRTs), this depiction does not fully capture the collaborative and distributed nature of incident response in practice. According to Kleij *et al.* [14], the assumption that a single, autonomous, and continuously operational team can manage all aspects of incident response often overlooks the interdependencies, communication demands, and *ad hoc* collaborations that characterize real-world incidents.

Instead, effective IR frequently requires a broader, more flexible approach involving participants from across the organization. NIST SP 800-61r3 [13] highlights that modern incident response extends beyond the scope of dedicated incident handlers and involves a wide array of internal and external stakeholders. While incident handlers remain critical-performing functions such as detection, analysis, containment, and recovery-effective response increasingly requires coordinated participation from leadership, technology professionals, legal counsel, human resources, public affairs, and facilities management. These roles contribute in domain-specific ways, such as legal review, personnel management, media communication, or physical access. Additionally, incident response is often supported by third-party actors, including MSSPs, cloud providers, and law enforcement, under

a shared responsibility model that demands clear contractual delineation of duties and communication protocols. This expanded model reflects the operational complexity of contemporary incidents and the need for a federated, cross-functional approach to ensure timely, compliant, and effective response actions. Bhaskar *et al.*'s [20] view of a security team as "mainly an *ad hoc* group of company employees who are assembled together in the event of an emergency" aligns with the previous research and NIST's perspective. These characteristics align closely with the concept of "task teams", "*ad hoc* teams" or "temporary organizations" formed to accomplish non-continuous tasks [17].

Positioning incident response (IR) roles and responsibilities within the framework of *ad hoc*, multidisciplinary task teams enables researchers to draw on a wider body of literature to better understand the specific challenges these teams face in terms of coordination and rapid team formation. The distinction is not that task teams are inherently superior to regular, co-located teams who work together routinely by virtue of organizational structure, but rather that task teams are an operational necessity given the multi-layered and cross-functional nature of incident response.

3.2 Understanding Team Mental Models (TMMs)

Team Mental Models (TMMs) are shared cognitive structures that enable team members to understand their tasks, coordinate their actions, and adapt to dynamic situations effectively. In one of the foundational studies, Cannon-Bowers describe TMMs as "knowledge structures held by members of a team that enable them to form accurate explanations and expectations for the task, and, in turn, to coordinate their actions and adapt their behavior to demands of the task and other team members" [21], or more simply, ensuring members are "on the same page" regarding tasks and coordination [22]. These shared representations emerge through interaction and allow members to leverage structured knowledge for coordinated action [23], [24].

3.2.1 Dimensions and Properties of TMMs

Research suggests that Team Mental Models (TMMs) are multi-dimensional constructs encompassing several interrelated properties. The most commonly cited dimensions, following the work of Mathieu *et al.* [24], [25], include the **task TMM** and the **team TMM**. The task TMM refers to a shared understanding among team members regarding the nature of the task itself. This includes common agreement on goals, procedures, strategies, relevant environmental conditions, and the equipment or technology involved. In contrast, the team TMM pertains to a shared cognitive representation of how the team functions.

It encompasses mutual awareness of roles and responsibilities, communication patterns, norms of interaction, and the distribution of expertise-essentially, who knows what within the team.

Key properties commonly used to evaluate Team Mental Models (TMMs) include **similarity** and **accuracy**. Similarity refers to the extent to which team members' mental models are aligned or overlapping, reflecting a shared cognitive framework that facilitates coordination and mutual understanding [26]. Similarity denotes the extent to which team members' mental representations of task goals, procedures, and interdependencies overlap, thereby enabling them to anticipate one another's actions and coordinate with minimal explicit communication [24], [27]. Empirical and meta-analytic studies show that higher cognitive similarity is linked to faster responses, fewer coordination errors, and superior performance in complex, interdependent tasks [22], [28]. However, excessive overlap can become counterproductive: uniformly shared mental models may stifle divergent thinking, encourage premature convergence on suboptimal solutions, and foster the kind of confirmation bias associated with groupthink[29]–[31]. Accordingly, researchers caution that effective teams balance a common situational framework with the integration of member-specific knowledge to avoid the pitfalls of over-similarity [28]. Both van der Haar[32] and Hällgren[31] suggests that it is possible to decrease the impact of groupthink as long as the team is promoting critical thinking.

Accuracy, on the other hand, denotes how closely a team's shared mental model mirrors a validated reference-typically an expert map of task requirements, system state, or role interdependencies-and is commonly quantified through expert-scored knowledge tests or structural comparisons such as Pathfinder network matching [27], [33]. Scholars distinguish taskwork accuracy (facts, procedures, environmental cues) from teamwork accuracy (role understanding, coordination logic); both facets have shown positive, and sometimes independent, links to performance [26], [34]. Empirical evidence across domains supports its value: concept-mapping studies tie greater accuracy to faster problem detection in command-and-control simulations [27]. Yet accuracy is not a panacea. Establishing a single "gold standard" can be contentious in ambiguous, evolving environments, and static referents quickly lose relevance; teams whose models were initially accurate but not updated have drifted into critical misjudgments [30].

These two dimensions and properties interact as well. For instance, high similarity on an inaccurate model can be detrimental [24]. Both properties are foundational in assessing the effectiveness of TMMs in supporting team performance, especially in complex and high-stakes domains such as cybersecurity. To simplify, we can say similarity is a question of "Are we on the same page?" while accuracy is a question of "Are we on the correct

page?". There is a possibility for the team members to be "on the same page" (high similarity) but on the "wrong page" (low accuracy).

3.2.2 Team Processes and Team Effectiveness

Team processes are defined as "member's independent acts that convert inputs to outcomes through cognitive, verbal, and behavioral activities directed toward organizing taskwork to achieve collective goals" [35], where taskwork is described as "a team's interactions with tasks, tools, machines, and systems" [35], [36]. Taskwork refers to what the team must accomplish, whereas teamwork captures how members collaborate to accomplish it. High-quality taskwork is vital for overall performance and rests on both individual expertise and the team's coordinating routines. Those routines-team processes-serve to steer, synchronize, and track the work. Although the line between taskwork and teamwork often blurs in real settings, our emphasis here is on the coordination mechanisms that knit task activities together so the team reaches its objectives [35]. In team processes research, several behavioral and cognitive mechanisms have been monitored, and some of them are described in upcoming sections.

Implicit Coordination

Implicit coordination is a dynamic and emergent team process wherein members anticipate each other's needs and actions based on a shared mental model of the task environment, allowing them to adjust their own behavior accordingly without the need for explicit communication or pre-assigned roles [37]. This anticipatory adjustment is rooted in the team's collective understanding of roles, objectives, and situational cues, which enables members to synchronize actions and allocate tasks fluidly as conditions evolve. Unlike explicit coordination, which relies on structured communication and formal delegation, implicit coordination emerges through repeated interaction, mutual trust, and shared cognitive frameworks. It is particularly crucial in high-tempo environments such as incident response, where time constraints and information overload may render overt coordination impractical or even counterproductive.

Rico*et al.* [37] argue that implicit coordination becomes increasingly important as task complexity and interdependence rise, especially under uncertainty or stress. In such settings, successful team performance often hinges on members' ability to detect environmental cues, infer teammates' likely responses, and act in ways that complement those responses. This process requires a high degree of cognitive alignment, often facilitated by prior experience, cross-training, or interventions aimed at developing Team Mental Models (TMMs). Accordingly, implicit coordination is not merely a byproduct of team familiarity,

but a measurable and trainable capability that reflects deeper cognitive integration within the team.

Shared Situational Awareness

While there is no commonly accepted definition for situational awareness (SA), it is generally perceived as knowing what's going on, why it matters right now, and what is likely to happen next in relation to your objectives. Since SA is subjective [38], there are many definitions but most referenced definition belongs to Endsley:

> Situation awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future [39]

This definition contains four aspects of SA: *Perception* is the initial scan that gathers the observable facts; *Comprehension* links those raw observations to one's stored expertise, turning data into meaning; *Projection* extends that meaning forward, forming a mental picture of how events will probably unfold if nothing external intervenes; and *Prediction* layers on an extra step, judging how possible external influences-such as new information, actors, or environmental shifts-could alter that projected course [38]. This definition is very close to what a *mental model* is, and the distinction is made such that a mental model is the schema of a system while SA is a temporal state or schemata of a moment in time that can evolve in time [39]. This distinction gets blurry when the temporal aspect of TMMs are discussed. Situational awareness takes on a different meaning when it refers to a single person versus when it describes a team or group. To clarify, Nofi argues that achieving shared situational awareness depends on team members explicitly exchanging their individual mental models of the situation; through this communication, a common operating picture emerges. In short, effective communication is the decisive factor in building and maintaining shared situational awareness [38].

Team Adaptation

Team adaptation is related to team-level behavioral and cognitive changes against changing situations. Though there are different approaches to it. Maynard *et al.* [40] analyzes team adaptation in an input-process-mediator-output model, where *team adaptation* is defined as the adjustments to relevant *team adaptation processes*-action, interpersonal, and transition-made in response to a disruption or trigger, whereby inputs that reflect the *team adaptability* are converted into *team adaptive outcomes* through mediators such as communication, coordination, and cognition. The overall process is best defined in Figure

1 in detail. In this approach team adaptation, is not a part of TMMs, on the contrary, TMM is an antecedent of team adaptation. Since team effectiveness develops over time and spans several dimensions, according to Resick [41], outcomes like performance results, behavior such as work processes and team adaptability, and cognitive elements that include members' attitudes and perceptions, it is reasonable to connect the team effectiveness to TMM via adaptability.

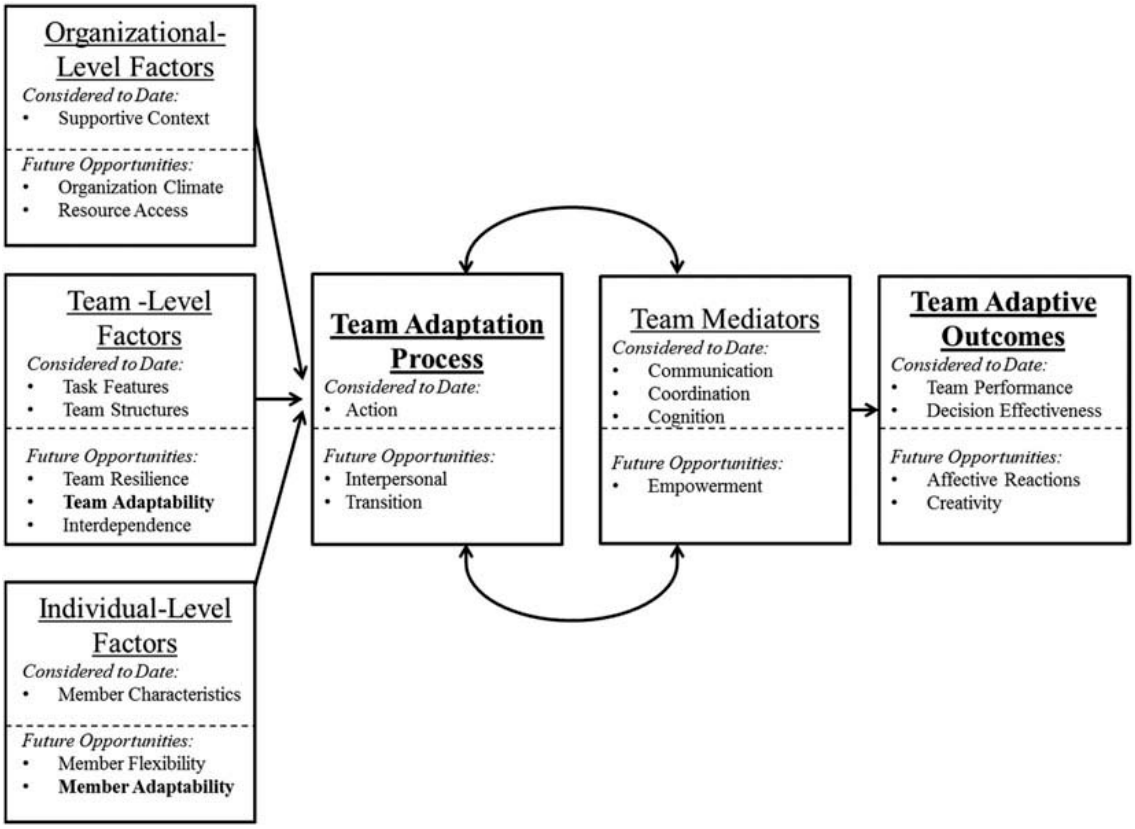

Figure 1. Input-throughput- output model of team adaptation. [40]

On the other hand, Burke *et al.* conceptualize team adaptation as an output within an input–process–output (IPO) model. Figure 2 illustrates this model, where both individual attributes and job-design features contribute to a team's adaptive capacity [42]. These individual attributes include task and team expertise, preexisting mental models, team-oriented attitudes, openness to experience, and cognitive ability. The job-design feature emphasized in the model is self-management. Together, these inputs provide the foundational resources for adaptation. When an environmental cue indicates a change, these inputs initiate a four-phase adaptive cycle. In the first phase, the team assesses the situation by detecting cues and assigning meaning to them. Next, they formulate a response plan. The third phase involves executing the plan through mutual monitoring, communication, back-up behavior, and leadership. Finally, the team engages in reflective learning before the next cycle begins. Each phase is flanked by time-sensitive emergent states, such as shared

mental models, team situational awareness, and psychological safety. According to Burke [42], these states influence—and are influenced by—the team's ongoing actions, forming recursive links between cognition, emotion, and behavior. When the cycle is successfully completed, it results in team adaptation. This adaptation can manifest as innovation or behavioral modification and contributes to overall adaptive team performance. A feedback loop then returns insights from this cycle back to the input side, preparing the team for future challenges. Both Team Mental Models (referred to in the model as shared mental models) and shared situational awareness (referred to as team situational awareness) are classified as emergent states. These are defined as "constructs that characterize properties of the team that are typically dynamic in nature and vary as a function of team context, inputs, processes, and outcomes" [43].

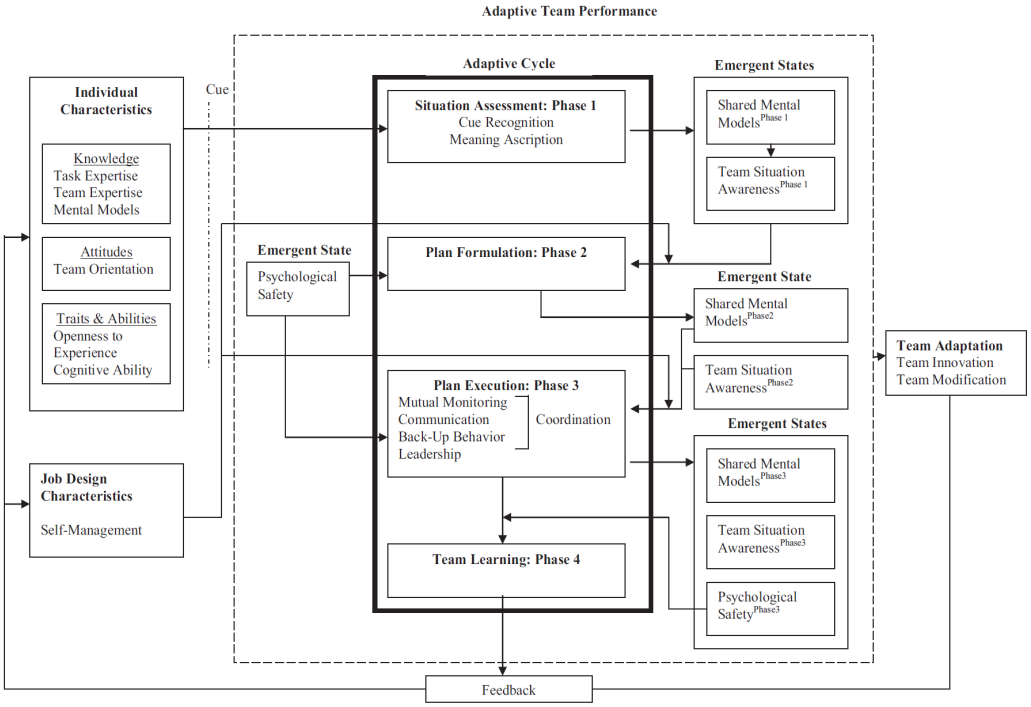

Figure 2. Team adaptation nomological network. [42]

In both models, team adaptation is considered a property of team effectiveness and TMM is depicted as a crucial component of team adaptation, either as an antecedent or an enabler in multi-level feedback loop.

Big Five

The article "Is There a 'Big Five' in Teamwork?" by Salas *et al.* [44] identifies five core components essential for effective teamwork. According to the authors, the word core is selected to express that there are other variables affecting the teamwork. The first of these is **Team Leadership**, which encompasses the direction and coordination of

team activities, performance assessment, skill development, motivation, and the creation of a positive team environment. Effective leadership clarifies roles and fosters support within the team. Secondly, **Mutual Performance Monitoring** is crucial, referring to the capacity of team members to observe each other's performance, offer feedback, and provide assistance. This relies on a shared understanding of team goals and individual contributions, enabling timely adjustments and mutual support. The third component, **Backup Behavior**, involves team members actively assisting one another, particularly during high-demand situations or when a colleague faces difficulties, by anticipating needs and offering support to sustain overall team performance. Following this is **Adaptability**, defined as the team's proficiency in modifying its strategies and actions in response to new information or evolving circumstances, which includes recognizing ineffective approaches and collectively transitioning to better alternatives, mentioned in Team Adaptation section above. Finally, **Team Orientation** signifies the collective inclination of members to prioritize team objectives over individual ambitions, fostering cohesion, a shared sense of purpose, and a strong belief in the team's mission.

The same research defines 3 coordinating mechanisms: shared mental models[1], closed-loop communication, and mutual trust. The Figure 3 shows the relationship of TMM, being part of the core and a support element for mutual performance monitoring, back-up behavior and adaptability components.

[1]For the sake of this research shared mental model and team mental model terms are used interchangeably.

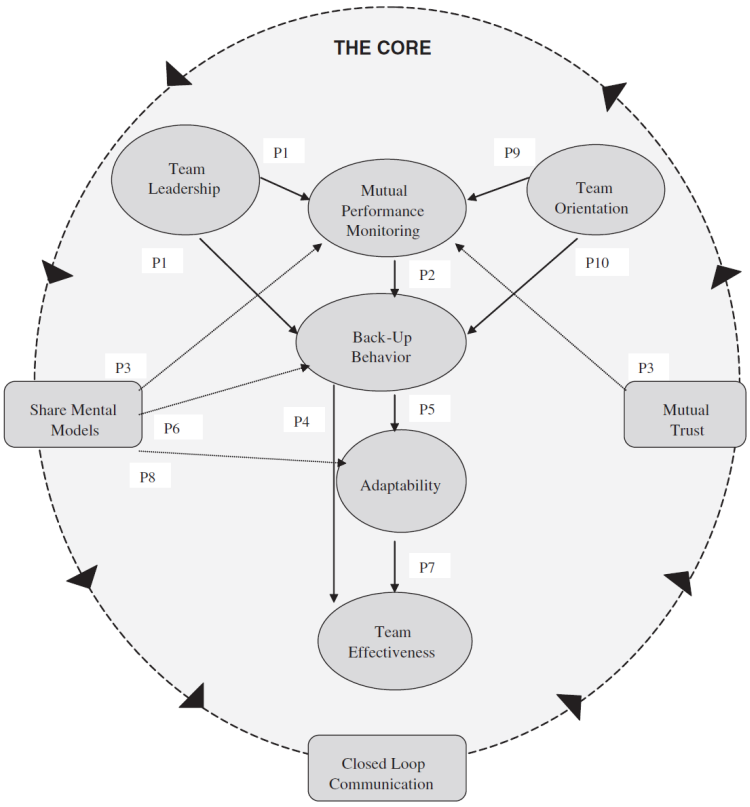

Figure 3. Graphical Representation of High-Level Relationship Among the Big Five and the Coordinating Mechanisms Including Research Propositions. [44]

While Salas *et al.* describe the Big Five as "core teamwork dimensions," they also argue that these dimensions provide a bridge between academic theory and practical application by offering a usable framework for understanding team processes [44]. The authors suggest that the Big Five framework enhances our understanding of how team processes evolve, particularly in relation to a team's ability to perform core coordination and communication tasks. In contrast, recent literature tends to interpret the Big Five not as dimensions but as "team processes" in their own right [45]. Figure 4 illustrates how these components have been reframed as primary team processes, with coordinating mechanisms positioned as supporting elements. Although the study by Johnsen *et al.* does not explain why this conceptual shift was made, the change appears to support their goal of making the Big Five constructs more easily measurable.

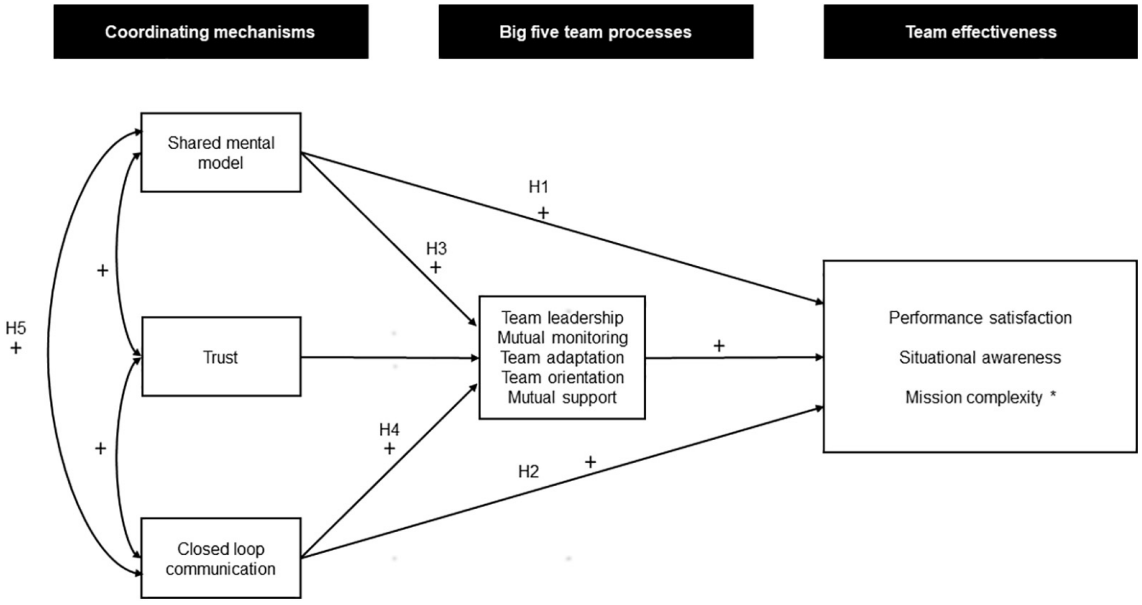

Figure 4. Proposed model of Johnsen *et al.* considering "Big Five" components as team processes
Source: [45]

Team Effectiveness

It is seen that studies regarding team effectiveness have different approaches. While closed loop communication and TMMs are considered as coordinating mechanisms improving teamwork, meaning processes in Big Five, and indirectly the team effectiveness [44], some research in dispersed teams suggests these coordinating mechanisms -TMMs and structured communication- might have a more direct impact on effectiveness than the mediating processes themselves [24], [45]. When it comes to similarity and accuracy aspects, Lim and Klein [34] reported a direct positive relationship between TMM similarity and team performance in their field study. A more recent and broader meta-analysis by DeChurch and Mesmer-Magnus [46] confirms a strong positive relationship between TMMs (both similarity and accuracy) and team performance, mediated by team processes. However, other studies suggest interactions are key; Smith-Jentsch *et al.* [47] found that neither task nor team TMMs alone directly affected efficiency, but their interaction did.

In either way, the results suggest direct and/or indirect impact of TMMs on team effectiveness.

3.3 Team Mental Models in Task Teams: Challenges and Imperatives

Across disciplines, a range of terms—such as "task teams," "ad hoc teams," "temporary organizations," "fluid teams," and "rapid reaction teams"—are used to describe similar

concepts: small, multidisciplinary groups of experts assembled for a specific purpose rather than working together on a daily basis. This study draws on all of these labels, synthesizing insights from their respective literatures to build a unified understanding of such *ad hoc* team structures.

In commercial aviation, the requirement comes from the fact that teams do not operate together continuously; instead, their composition typically changes with nearly every flight. These teams can thus be considered *ad hoc* or swift-starting teams [48], which places additional demands on rapid team-building and the quick establishment of shared understanding and coordination [49]. The accidents caused by problems in human communication and coordination and the solution was implementing a teamwork and leadership focused training method called Crew Resource Management (CRM) [50] [49].

Crew Resource Management (CRM) originated in the aftermath of several high-profile aviation accidents in the 1970s that were attributed not to technical malfunctions but to breakdowns in teamwork, leadership, and communication [51]. In response, NASA held a pivotal workshop in 1979 that highlighted the role of interpersonal and cognitive factors in cockpit errors, marking the formal beginning of CRM training. The initial programs, termed "Cockpit Resource Management," focused on individual behaviors such as assertiveness and leadership style. Over time, CRM evolved through multiple generations: from psychological seminars to team-based modules, and eventually to integrated training embedded in operational and technical procedures. By the 1990s, CRM was broadened to include all flight-related personnel, and later reframed as a comprehensive strategy for error management that included avoiding, trapping, and mitigating human errors. Table 1 is adapted from historical evolution of CRM approach to depict both trends and contribution to teams.

Table 1. CRM Skills by Generation. Adapted from [51]

CRM Generation	Skill	Notes
First Generation	Leadership	Addressed due to concerns about authoritarian captains and insufficient crew input.
	Assertiveness	Encouraged junior crew to challenge decisions when appropriate, based on accident analysis.
	Communication	Initial programs emphasized cockpit interpersonal dynamics.
Second Generation	**Team Building**	Introduced to improve cockpit group dynamics.
	Briefing Strategies	Promoted shared understanding in temporary crew compositions.
	Decision-Making	Structured approaches introduced to break error chains.
	Situational Awareness	Trained crews to maintain current and predictive understanding of the environment.
	Stress Management	Addressed cognitive limitations and performance degradation under stress.
Third Generation	Leadership	Specialized modules developed for captains assuming command roles.
	Use of Cockpit Automation	Focused on managing flight management systems and reverting to manual control under pressure.
Fourth Generation	Proceduralized Behaviors	CRM practices integrated into SOPs and checklists for consistency.
Fifth Generation	Error Management (Avoid, Trap, Mitigate)	CRM reframed as structured behavioral countermeasures to manage human error.

Note: Emphasis on skills are added to emphasize the alignment of CRM research with TMM related topics.

In emergency and disaster response area, there are cross-agency teams involved in responding to natural disasters and regional or national crisis situations. As per van der Haar's research [32], emergency management teams consist of highly skilled individuals collaborating to carry out urgent, unpredictable, interdependent, and high-stakes tasks, even while team composition frequently changes. These teams share a clear, common goal and bring together diverse expertise and resources to tackle tasks that inherently require coordinated teamwork. Due to the frequent shifts in team membership, these teams often lack a prior history of working together. Thus, they must rapidly learn effective collaboration methods suited to their current team makeup and the specific demands of each unique crisis situation. Table 2 summarizes the key characteristics of *ad hoc* multidisciplinary emergency management teams, compared to regular teams. While van der Haar outlines various features of task teams, only some are unique to their *ad hoc* nature. Both team types share traits like high skill, interdependence, and teamwork under a common goal. However, task teams differ in operating under time pressure, unpredictability, and frequent changes in composition-features less typical of regular teams with more stable membership.

Table 2. Comparison of Task Teams and Regular Teams Based on Core Characteristics. Adapted from [32]

Category	Characteristic	Task Teams	Regular Teams
Team	Highly skilled members	+	+
	Mix of experience and resources	+	+
	Interdependent	+	+
	Frequent changes in composition	+	-
Task	**Urgent, immediate response needed**	+	-
	Unpredictable	+	-
	Highly consequential	+	+
	Clear common goal	+	+
	Training novice members	+	+
	Requires teamwork	+	+

Note: Emphasis and regular teams comparison is not in the original research.

A similar research on task teams is conducted under the term "fluid teams". Fluid teams are defined as "teams that are rapidly assembled from across disciplines or areas of expertise to address a near-term problem" [52]. Fluid teams -in our case, *ad hoc* teams or task teams- differ in several key ways from traditional teams, according to Driskell [52], [53]. They are rapidly assembled, often bringing together members from diverse disciplines, which typically results in limited prior familiarity among team members. Given the urgent nature of their tasks, these teams have little time to orient themselves or build rapport. Operating within short time frames-ranging from a few hours to several days-fluid teams lack the opportunity to gradually develop team cohesion or shared mental models. Furthermore, they dissolve immediately upon task completion, with no expectation of future collaboration. These characteristics create distinct challenges for effective team performance. To address these, the source proposes a research agenda focused on selection, design, and training strategies tailored to the needs of fluid teams: *leadership training*, *trust building*, *pre-briefing activities*, *role definition and clarification*. Fluid teams research does mention shared mental models as one of the targets in team training however does not provide empirical data and notes this as a topic needing further research.

In addition to the aforementioned areas of research, *ad hoc* teams are studied in medicine, especially emergency rooms. In hospital medicine, teams are often *ad hoc*, meaning their composition frequently changes. This is especially common in academic teaching hospitals, where team members rotate across shifts and training schedules. The fluctuating membership presents challenges, as these teams often lack the time and continuity needed to build a shared identity, develop mutual trust, or establish common mental models [54]. In emergency rooms, teams carry two distinct characteristics. First, they operate under

severe time constraints, which limits opportunities for detailed planning and structured communication during care delivery. Second, these teams are typically *ad hoc* in nature, with members frequently working in shifting compositions. For such teams to function effectively, it is essential that roles and responsibilities are clearly defined and that communication and leadership are well coordinated. However, these conditions are not always met in practice. For example, research has shown that two-thirds of rapid response teams failed to restore normal heart rhythm in cardiac arrest cases, largely due to deficiencies in communication and leadership [55]. Hence, the training efforts are focusing on leadership, team communication, and role-appropriate team behaviors [55]. This aligns with similar approach as task team research in medicine also tries to utilize CRM as a tool [56] imported from aviation.

In a study focused on applications of team mental models in healthcare, Burtscher *et al.* [57] identify two organizational configurations where TMM research holds particular relevance: inter-professional teams and multi-team systems. According to their analysis, inter-professional teams-comprising consultants, junior doctors, nurses, cardiac technicians, physiotherapists, and other healthcare providers-operate in diverse clinical contexts ranging from intensive care to psychiatric rehabilitation. The heterogeneity of these teams presents a challenge for developing uniform approaches to TMM assessment. Nevertheless, Burtscher *et al.* argue that TMM theory provides a valuable framework for addressing professional friction arising from divergent views on patient care. They suggest that promoting a shared understanding of patient-centered priorities could mitigate inter-professional conflict and enhance team effectiveness. In this context, the authors propose that attitudinal components of TMMs, such as mutual conceptions of effective communication and a shared commitment to patient safety, may offer a more feasible focus than procedural consensus. They further posit that the similarity of team members' mental models-rather than their objective accuracy-is a more appropriate metric in such settings, and that interventions such as cross-training could play a key role in developing aligned understandings of teamwork.

In the same work, Burtscher *et al.* also explore the role of TMMs in multi-team systems, particularly within operating room environments. Here, they describe how clinical staff simultaneously function as an overarching unit and as separate sub-teams, typically comprising surgeons, anesthetists, and nurses. These sub-teams alternate between phases of close integration and periods of independent operation. The authors suggest that this structural dynamic creates the need for dual-level mental models: one that is task-specific within each sub-team and another that supports coordination at the inter-team level. Although Burtscher *et al.* note that empirical research in this area is still emerging, they contend that existing overviews of TMM measurement methods offer a solid foundation for the

development of future studies in multi-team clinical contexts.

This study focuses on task teams even named differently: short-lived, purpose-driven, and cross-functional in nature. Among the two team types discussed by Burtscher *et al.* [57], inter-professional teams align with our scope due to their dynamic composition and need for rapid coordination. In contrast, multi-team systems, which involve more stable sub-team structures and recurring collaboration patterns, fall outside the focus of this work.

It has been observed in academic literature that different research domains employ varying taxonomies to describe identical or closely related concepts. Although the development of a universal theory of team research remains challenging, teamwork continues to be approached as a multidisciplinary subject. This diversity in terminological frameworks contributes to conceptual fragmentation. The absence of a common taxonomy is believed to hinder the formation of shared mental models concerning teamwork, team processes, and, paradoxically, team mental models themselves.

Incident-response (IR) teams stand to benefit from the section's findings because they display the very characteristics—rapid assembly, heterogeneous expertise, high time-pressure, unpredictable task trajectories, and immediate dissolution—examined in aviation, disaster management, and emergency medicine research on task, ad-hoc, and fluid teams; empirical evidence from those domains shows that cultivating shared Team Mental Models (TMMs) through structured briefings, clear role definition, and checklist-driven coordination measurably increases decision accuracy, reduces error cascades, and shortens resolution times, suggesting that IR teams can reap parallel gains in faster containment, lower mean-time-to-recover, and improved cross-functional collaboration by adopting the same cognitive and communication scaffolds.

3.4 Strategies for Developing Team Mental Models

Ad hoc incident-response teams typically lack the pre-existing familiarity and shared experiences that drive organic development of team mental models (TMMs), making rapid cognitive alignment both critical and difficult (see Section 3.3). To overcome this, researchers and practitioners have devised targeted interventions that deliberately scaffold shared cognition. Cross-training rotates team members through each other's roles, building inter-positional knowledge and yielding a small-to-moderate performance gain ($r \approx 0.29$) in meta-analytic studies [58]. Pre-task strategy briefings or Team Dimensional Training focus team discussion on goals, roles and contingencies, and have been shown to produce significant improvements in mental-model accuracy and flexibility across routine and novel tasks [59]. Simulation-based exercises immerse teams in realistic scenarios with built-in

feedback, fostering both perceptual cue uptake and coordination skills—studies in military and clinical domains report 20–25% performance gains from brief debriefings alone [60]. Finally, structured guided self-correction (debriefing) sessions elicit reflection on errors and successes, driving up to a 25% increase in learning outcomes and medium-to-large gains in teamwork processes after just one session [60], [61]. Together, these techniques accelerate the convergence of TMM similarity and accuracy by aligning individual understandings around task, equipment and interaction schemas —an essential capability when teams must "get on the same page" and "get on the correct page" under severe pressure.

Cross-Training

Cross-training involves rotating team members through each other's roles so that individuals learn not only their own responsibilities but also the tasks, equipment use and decision criteria of their peers. This approach is grounded in the Shared Mental Model (SMM) framework, which identifies four key types of shared cognition—equipment, task, interaction, and team—each of which supports coordination and performance [62]. Moreover, cross-training develops inter-positional knowledge (IPK) —a term coined by Volpe *et al.* [63] described as "a type of role knowledge in which team members have information regarding the appropriate system operation behavior for each interdependent member within the team structure", and was used prior to the widespread adoption of the term *shared mental models*- by focusing explicitly on the roles and responsibilities of other team members [62], [63]. By fostering a more accurate and consistent awareness of each member's role, cross-training enables individuals to formulate a shared perception of the task environment[64]. Empirical studies underscore cross-training's value: naval cadet teams showed significantly higher performance after a single cross-training session compared to controls [65], and cross-training has been shown to improve team-interaction mental models [66] as well as more accurate and shared taskwork and teamwork models[30]. A meta-analysis by Salas, Nichols, and Driskell (2007) found that, although cross-training's unique contribution to performance ($r = .289,\ z = 0.544,\ p = .293$) was not significant when isolated, team-training interventions that combine cross-training with adaptive coordination and guided self-correction yield stronger gains overall ($r = .29$)[58]. These findings suggest that the effectiveness of cross-training depends on clearly defining the target SMM type ("What")—for example, whether the focus is on task procedures or interpersonal interaction patterns—and on selecting complementary strategies ("How") such as debriefs or simulation exercises to reinforce the newly acquired shared cognitions [66]. While potentially resource-intensive, even brief cross-training or targeted role-familiarization drills may therefore yield substantial benefits for *ad hoc* teams when embedded in a broader, multi-component training regimen.

Simulation-Based Training (SBT)

Simulation-Based Training places teams in realistic, dynamic scenarios that mirror the demands and constraints of their actual tasks, thereby enabling shared experiential learning and the implicit construction of both *task* and *team TMMs* through collaborative problem solving [67], [68]. In domains where live practice is too costly or hazardous—such as cybersecurity incident response—task simulations provide rich contextual instruction that fosters shared strategic knowledge while preserving safety [69]. Effective SBT for team situational awareness must incorporate scenarios that challenge individuals to assess critical cues and that require coordinated team processes, from rapid information exchange to decision-making under pressure [69]. Moreover, feedback loops are essential: participants need timely, specific feedback on how their actions and interactions influence both individual performance and team outcomes. A thorough task-analysis should precede scenario design to identify the most consequential and difficult contingencies that teams are likely to face, ensuring that practice is targeted to those high-impact events. The tabletop exercises employed in this thesis adhere to these principles by offering a safe yet demanding environment in which *ad hoc* IR teams can rehearse cue recognition, interaction protocols, and adaptive coordination—all critical to forging robust shared mental models before confronting real-world cyber incidents.

Team Dimensional Training / Strategy Briefings

This approach focuses on explicit discussion and clarification of team goals, strategies, roles, expectations and potential contingencies before task execution, with the aim of establishing a shared baseline of task and team TMMs that members can draw upon during performance [59], [61], [62]. Empirical work by Marks and colleagues demonstrated that both structured team-interaction training and enhanced leader briefings yield significant main effects on mental-model similarity and accuracy across routine and novel environments [59]. These interventions not only prepared teams to confront varied scenarios by fostering accurate and similar knowledge structures, but also improved the flexibility of member mental models, enabling coordinated adaptation when circumstances changed. Critically, enhanced briefings and interaction training were associated with more effective communication processes and higher overall team performance, with the linkage between shared mental models and performance being particularly pronounced in novel, high-uncertainty contexts. For *ad hoc* teams facing time constraints, such pre-task briefings offer a time-efficient means of achieving initial cognitive alignment and equipping members to adjust their shared cognitions as the task environment evolves.

Guided Team Self-Correction / Debriefing

Guided Team Self-Correction, commonly known as debriefing, refers to a structured, expert-facilitated review conducted immediately after task execution. During this process, teams systematically reflect on their actions, identify coordination breakdowns and set concrete improvement goals guided by a cognitive model of ideal performance [61]. Smith-Jentsch *et al.*'s empirical work in a military context demonstrated that a single guided debrief significantly enhances team-mental-model accuracy and teamwork processes compared to conventional reviews [61]. Meta-analytic evidence across domains such as healthcare, aviation and first responders shows that debriefings under twenty minutes yield average performance gains of 20–25% and medium-to-large effects on learning and non-technical skills [60]. Cybersecurity incident response shares the high-stakes, time-pressured and interdependent task structure of these settings, making brief, facilitator-led debriefs directly applicable to IR training. In our tabletop exercises, inserting a five- to ten-minute guided debrief immediately after each scenario not only reinforces the micro-intervention's focus on shared perception of critical cues but also provides the reflective feedback loop needed to translate individual insights into a coherent, collective understanding. This integration of debriefing into the TMM protocol leverages proven military and clinical practices to accelerate the development and retention of robust shared mental models in *ad hoc* teams facing evolving cyber threats.

Storytelling

Storytelling, according to Tesler *et al.*, involves the planned use of narrative as a structured team intervention, aiming to increase the similarity and accuracy of team mental models (TMMs) by providing memorable and relatable scenarios. Unlike purely analytical briefings or structured reflection sessions, storytelling explicitly leverages emotional and cognitive engagement through narrative structure—presenting a coherent story with clear beginning, conflict, and resolution—to facilitate shared understanding and retention of core teamwork concepts such as timing, communication, and coordination [70]. Tesler *et al.* empirically demonstrated that storytelling significantly enhanced TMM similarity when combined with guided team reflexivity sessions. Specifically, storytelling improved team members' shared mental models about task sequence and coordination requirements, resulting in measurable performance benefits compared to groups who received identical factual content without narrative context. Thus, storytelling represents a viable complementary training approach, reinforcing analytical briefing and debriefing strategies by anchoring cognitive lessons within emotionally salient, memorable scenarios.

Communication Protocols and Tools

Structured communication protocols—most notably Closed-Loop Communication (CLC)—and standardized reporting formats serve as coordinating mechanisms that underpin both the "Big Five" team processes and the continuous updating of shared mental models [44]. Empirical work by Johnsen *et al.* in geographically dispersed emergency-dispatch teams demonstrated that CLC exerts one of the strongest influences on team processes (unstandardized $\beta \approx 3.19$, $p < .001$), which in turn supports performance satisfaction, situational awareness and interrelated Shared Mental Models (SMMs) [45]. These findings indicate that confirmatory exchanges—where senders seek explicit acknowledgment of received information—both reduce misunderstandings and furnish the feedback loops essential for revising task- and team-related mental models, particularly when direct observation is impossible.

Meta-analytic evidence further confirms the power of structured communication: Salas *et al.* (2005) reviewed multiple studies and found that protocols like CLC and standardized checklists yield medium to large correlations ($r \approx 0.30$–0.45) with measures of coordination and overall team performance [44]. Technology platforms—such as shared incident-management dashboards and live incident timelines—extend these principles by creating a persistent, common operational picture that reinforces equipment and interaction SMMs in real time. In our tabletop IR exercises, embedding CLC expectations into every scenario and using a shared digital whiteboard will ensure participants continually articulate, confirm and correct their mental representations of both technical events and each other's roles, thereby strengthening both the similarity and accuracy of team mental models under pressure.

3.5 Measurement Considerations and Critical Perspectives on TMM Research

While the Team Mental Model (TMM) construct offers valuable insights into team cognition and performance, a critical perspective acknowledges several ongoing debates, limitations, and measurement challenges within the literature. Understanding these nuances is crucial for interpreting TMM research and applying the concepts appropriately, particularly in complex domains like cybersecurity incident response.

3.5.1 Heterogeneity in TMM Measurement

A significant challenge in synthesizing TMM research is the lack of standardized measurement approaches. The meta-analysis by Burtscher *et al.* [57] highlighted this heterogeneity across 33 studies, revealing diverse methods for assessing different TMM types and properties. Table 3 provides a summary adapted from their findings.

Table 3. Summary of TMM Measurement Approaches in Selected Studies (Adapted from [57])

Cluster	Common TMM Types (Studies)	Property (Studies)	Common Measurement Methods (Studies)	Common Analysis Methods (Studies)
Student Project Teams	Team (3), Attitudes (2)	Similarity (5)	Likert questionnaire (2), Open-ended Q (2)	rwg (2), Content analysis (1), Higgins Formula (1)
Command & Control Teams	Task (7), Team (3)	Similarity (10)	Pair-wise comparison ratings (7), Concept Map (3)	Pathfinder (4), Indiv. Coding (3), UCINET (2)
	Interaction (3)	Accuracy (5)		
Negotiation Teams	Task (1), Team (1)	Similarity (4)	Likert questionnaire (3), TMM as IV (2)	Avg. Agreement (3), Content analysis (1)
Business & Service Teams	Attitudes (2), Team (2)	Similarity (5)	Likert questionnaire (3), Pair-wise ratings (2)	Coeff. Variation (1), Avg. Agreement (1), MDS (1)
Action Teams	Task (4), Interaction (2)	Similarity (6)	Card sorting (3), Likert questionnaire (2)	Phi coeff (2), Pathfinder (1), Avg. Corr (1)
	Team (1)	Accuracy (5)	Pair-wise comparison ratings (1)	Avg. Euclidean Dist (1), rwg (1)
Overall (Multiple Mentions)	**Task (13), Team (11)**	**Accuracy (9)**	**Likert Q (10), Pair-wise ratings (9)**	**Pathfinder (5), Avg. Agreement (4), rwg (3)**
	Interaction (7)		**Concept Map (3), Open-ended Q (3)**	**Indiv. Coding (3), Content Analysis (2)**

Note: Numbers in parentheses indicate the count of studies within that cluster/category using the specified type/property/method/analysis in the Burtscher *et al.* (2012) review. This table simplifies the original for clarity.

As summarized in Table 3, the meta-analysis conducted by Burtscher *et al.* [57] illustrates the considerable methodological heterogeneity in how team mental models (TMMs) have been studied across domains. The table categorizes 33 empirical studies by team cluster, TMM type, measured property, data collection methods, and analytic techniques. In terms of TMM types, the most frequently examined categories were task models (appearing in 13 studies) and team models (11 studies), followed by interaction models (7 studies) and attitude-based models (4 studies). These types were not evenly distributed across domains; for example, task TMMs were particularly common in command and control teams, while attitude and team models appeared more often in student, business, and service teams.

Regarding TMM properties, similarity was by far the most frequently assessed attribute across clusters. It appeared in the majority of studies regardless of team type, including negotiation, action, and project teams. In contrast, accuracy was assessed in only 9 studies overall and was primarily associated with command and control and action teams. Complexity, although conceptually distinct from similarity and accuracy, was not explicitly addressed in any of the studies included in the review.

Measurement methods also varied considerably. Likert-scale questionnaires were the most commonly used technique, appearing in 10 studies across various clusters. Pairwise comparison ratings were used in 9 studies, particularly in command and control and business settings. Other approaches such as concept mapping, card sorting, and open-ended questions were present but less frequent. For example, concept maps were used in 3 studies, and open-ended responses were reported in another 3. Card sorting appeared mainly in action teams.

The analytic methods were equally diverse. Pathfinder, a structural analysis tool, was used in 5 studies-mostly in the command and control and action team domains. Average agreement methods, such as interrater agreement indices, were used in 4 studies. The *rwg* statistic, which measures within-group agreement, appeared in 3 studies. Other analytical approaches included individual coding (3 studies), content analysis (2 studies), multidimensional scaling (1 study), coefficient of variation (1 study), and Euclidean distance (1 study). Some studies employed multiple techniques, and several analytic methods were applied to both quantitative (e.g., Likert data) and qualitative (e.g., open-ended responses) sources.

Across clusters, different combinations of TMM types, properties, and methods were used, suggesting that choices may have been shaped by contextual or domain-specific factors. However, no single dominant approach was observed. The wide range of data collection and analysis strategies across the studies in this review demonstrates the diversity

of methodological choices in the TMM literature up to that point. This variation is not limited to individual studies but reflects broader patterns across team types and settings, highlighting the fragmented state of measurement practices in the field. While this review does not prescribe a unified method, it systematically documents the approaches used and presents a clear picture of the field's methodological diversity.

3.5.2 Ongoing Debates and Critical Considerations

Beyond specific measurement techniques, several broader debates and limitations persist in the study of team mental models (TMMs). One key issue concerns the trade-offs between structural and perceptual measures, according to Mohammed *et al.* [22], [46], [71]. Structural methods, such as Pathfinder and concept mapping, offer objective representations but are often laborious to implement. In contrast, perceptual measures-including questionnaires and agreement indices like *rwg*-are more practical but introduce subjectivity [22], [46], [71].

Another ongoing debate involves the tension between context-specificity and generalizability. While there is a need for measures that are sensitive to the nuances of specific tasks, researchers also strive for instruments that can be applied across diverse teams and domains [71]. Relatedly, questions arise regarding the generalizability of findings derived from controlled laboratory settings or stable, experienced teams-such as those in aviation-to high-stress, dynamic, and unfamiliar environments like those faced by *ad hoc* incident response teams [72].

Further, Janis [29] points out that there are potential downsides to high levels of cognitive similarity or convergence within teams. Although such alignment may facilitate coordination, it can also suppress creativity, reduce vigilance, and increase susceptibility to groupthink, ultimately hindering the team's ability to respond effectively to novel situations. This underscores the importance of considering cognitive diversity and the role of constructive disagreement.

In addition, TMMs are dynamic and evolve over time through interaction and shared experiences [22], [73]. Yet many studies continue to rely on static pre/post measures, limiting our understanding of TMM development trajectories. There is a clear need for longitudinal research designs to better capture these temporal dynamics.

Finally, TMMs represent just one component within the broader system of team cognition. Constructs such as transactive memory systems-knowing who knows what-and shared situation awareness-understanding the current operational context-interact with TMMs

in complex ways. Measuring TMMs in isolation may therefore provide an incomplete or distorted picture of team cognitive functioning [67].

3.6 Team Mental Models, Task Teams and Cybersecurity Incident Response

3.6.1 TMM literature and IR

There are some recent work on mental models in cybersecurity as well. In one of the studies, Van den Berg offers an "essential set" of mental models designed to give managers, technicians, and policy-makers a shared vocabulary for analyzing cyber risk [74]. The framework begins with a three-layer view of cyberspace that distinguishes the technical IT/OT infrastructure, the socio-technical layer of human-system activity, and an outer governance layer of rules and institutions. Each layer is paired with a clarifying concept: the OSI/TCP-IP stack (augmented by the IT-OT split) for infrastructure, the four-stage learning ladder for secure behavior in the socio-technical realm, and Lessig's modalities of regulation-laws, norms, markets, and architecture-for governance. In the research, security is cast as a risk-management loop. Van den Berg adopts the ISO/TC 262 cycle (from critical-activity identification through monitoring) and couples it with two visual aids: the bow-tie model, which links threats, controls, and recovery measures, and the likelihood-by-impact risk matrix used for quick appraisal. The classical "avoid, transfer, accept, mitigate" response set shows how decisions flow from the matrix, while a Swiss-cheese depiction of defense-in-depth illustrates the value of layered controls. Finally, two governance-oriented diagrams-the institutionalization ladder for public-private partnerships and a direct-versus-indirect social-contract view of responsibilities-round out the toolkit. Together, these models form a concise repertoire that stakeholders can assemble as needed to reason about cyber threats and controls, although Van den Berg notes that detailed guidance on balanced mitigation selection and large-scale cooperation remains an open research priority.

In another work in mental models, Murimi and colleagues provide a decade-long review of how users conceptualize cybersecurity through mental models [75]. Drawing on more than forty studies, they divide existing work into two overarching categories: *folk models*, which rely on everyday analogies such as physical burglary, medical infection, criminal activity, warfare, or market failure, and *formal models*, which originate in engineering or cognitive science and include constructs such as error-and-blocking state diagrams, control-loop representations (e.g. OODA), firewall state models, encryption "black-box" schemata, and usable-security frameworks. Their survey shows that folk metaphors dominate popular risk communication, whereas formal models appear mainly in expert training and tool

design. The authors also report empirical evidence that users with formal cybersecurity exposure articulate longer, more domain-specific descriptions of threats than those with only informal exposure, indicating that expertise shapes both the richness and precision of mental model content. Across the surveyed literature, Murimi *et al.* [75] emphasize that mental models are neither universally "correct" nor "incorrect"; rather, their value lies in the behaviors they elicit. In some cases even incomplete or inaccurate conceptions can still prompt sound protective actions, whereas technically accurate models may fail to influence behavior at all. Consequently, the review argues that effective security interventions should focus less on teaching canonical models and more on fostering representations that reliably trigger desirable practices. To that end, the paper evaluates three broad design strategies: removing users from security decision loops ("stupid" approach), educating users in technical detail, and aligning interfaces and messages with the ways users already think about security ("understand-how-users-think" approach); the authors favor the latter two while warning against one-size-fits-all automation. The article closes by calling for a curated, standardized repository of widely used cybersecurity mental models-akin to the MITRE CVE database-to guide future research, training, and tool development, and it highlights the influence of social factors, workforce needs, and emerging technologies on the evolution of user mental models.

Empirical work that focuses directly on mental models within cybersecurity incident-response (IR) teams remains limited but illustrative. In one of the earliest large-scale studies, Tjaden and colleagues asked ninety practitioners from different Computer Security Incident Response Teams (CSIRTs) to enumerate the information they would need during a simulated botnet crisis; the resulting lists showed little overlap, leading the authors to conclude that the group “did not exhibit a shared mental model for decision making” [76]. Follow-up analyses ranked each information item by perceived importance and collection difficulty, producing a prioritized schema that could serve as a baseline expert model for training. Steinke *et al.* synthesized behavioral findings from nuclear-power control rooms, military command posts, and emergency medical services, mapping validated interventions-cross-training, guided self-correction, after-action reviews-onto CSIRT workflows [77]. Complementing these organizational studies, Maier’s work on expert-novice differences in cyber-attack visualization shows that dashboards aligned with expert mental representations reduce triage time for all user categories [78]. Kullman and co-authors extend this line by proposing stereoscopic visual analytics explicitly designed around analysts’ internal models of network evidence [79]. Despite such advances, the systematic measurement of TMM *accuracy*-as opposed to similarity-remains rare in Cyber-IR settings, echoing gaps documented in broader meta-analyses [57].

In cybersecurity area, mental models are lacking empirical studies in incident response in

general. Further research is needed.

3.6.2 Task Team literature and IR

As explained previously, what makes task teams different is that they have "frequent changes in composition" while their tasks are "urgent", "unpredictable", and requiring "immediate response" [32]. Being the unique characteristics, it is better to focus on this aspect.

Due to one of the characteristics frequent changes in team composition, task teams encounter challenges in maintaining stable TMMs. Driskell *et al.* argue that frequent membership changes significantly disrupt the similarity and stability of TMMs because each addition or removal of a member requires realignment of shared knowledge, roles, and expectations, hindering rapid coordination and cohesion [80]. This frequent recomposition is common in cyber incident response teams (CIRTs), which routinely integrate specialized personnel such as forensic experts or legal advisors, complicating the formation of stable and cohesive TMMs [13].

Additionally, urgency and the need for immediate response inherent in task teams place substantial demands on rapid cognitive synchronization among members. Lim and Klein observed that teams engaged in urgent tasks frequently rely on structured communication and predefined operating procedures to quickly align their mental models, which can result in sacrificing detailed accuracy checks for quicker action [34]. This urgency directly impacts CIRTs, where immediate responses to cyber threats such as ransomware or active intrusions are necessary. CIRTs commonly use structured incident-handling protocols like those outlined by NIST to expedite coordination and action under pressure [13].

Lastly, unpredictability significantly influences the accuracy and stability of TMMs within task teams. Uitdewilligen and Waller found that unpredictable task environments hinder a team's ability to maintain accurate and complete mental models, leading to continuous updates and adaptations in response to new information or evolving conditions [81]. This unpredictability is especially relevant to CIRTs, where cyber incidents constantly evolve. Hence, structured training exercises, such as scenario-based tabletop simulations recommended by NIST, become essential tools for improving adaptability and maintaining accurate mental models in cyber incident responders [82].

In addition to the focus common characteristics, cross-domain studies may provide examples of how task-team concepts translate into measurable outcomes. Aviation research shows that short, scripted briefings increase TMM similarity and reduce coordination

errors during abnormal flight procedures [83]. In emergency medicine, leadership assignment combined with a communication checklist significantly improves the likelihood of restoring sinus rhythm in cardiac-arrest cases, an effect attributed to faster alignment of team expectations [77]. Within cybersecurity, the exercise reported by Tjaden *et al.* demonstrated that even experienced responders may lack a common schema for prioritizing evidence, thereby prolonging investigation timelines [76]. Their proposed next steps-instrumented ticketing systems and virtual training environments designed around a shared incident schema-illustrate how domain-specific artifacts can serve as external memory aids that hasten TMM convergence. Steinke *et al.*'s comparative review further recommends cross-training and guided self-correction for CSIRTs, interventions already validated in military and healthcare teams [77]. Complementary conceptual work argues that a holistic set of mental models-spanning layers of cyberspace, risk-assessment cycles, and kill-chain stages-can give heterogeneous stakeholders a common vocabulary for discussing threats and mitigations [74]. Reassessing IR teams as task teams and making use of task team literature allows us to frame incident response in the teamwork perspective. However, it may be noted that considering MITRE ATT&CK Framework or the Cyber Kill Chain® by Lockheed Martin as mental model items is a valuable contribution in understanding TMM and cybersecurity.

3.6.3 Summary

The literature review provides a structured foundation for examining the potential influence of targeted training on Team Mental Models (TMM)—shared cognitive representations among team members regarding team roles, tasks, goals, and the operational environment—in improving the effectiveness of *ad hoc* task teams during cybersecurity incident response. *Ad hoc* task teams refer to temporary teams assembled rapidly to address specific tasks, typically with limited prior interaction. While direct research on TMM training within the specific context of incident response remains limited, insights were drawn from a broader set of disciplines where team-based decision-making under pressure is well studied. Domains such as aviation, military operations, emergency medicine, and high-reliability organizations have contributed substantially to the conceptual understanding of TMMs and their operational benefits. These fields provide robust empirical evidence that targeted training aimed at developing shared mental models can enhance team coordination, reduce miscommunication, and improve adaptive performance in dynamic, high-stakes environments.

Within the literature, a contrast is evident between traditional cybersecurity training—which emphasizes individual technical proficiency—and TMM-focused approaches, which aim to cultivate shared understanding across team members regarding roles, goals, and situational

dynamics. Notably, two key aspects of TMMs—similarity and accuracy—are consistently linked to improved team outcomes. Similarity refers to the degree of overlap or congruence among team members' mental representations ("Are we on the same page?"), while accuracy pertains to how well these shared representations match an expert-derived or objectively correct understanding of the task environment ("Are we on the correct page?"). There is a possibility for the team members to be "on the same page" (high similarity) but on the "wrong page" (low accuracy).

In addressing the first sub-research question, the review suggests that TMM-specific training leads to greater similarity in mental models among participants than conventional security training, thereby supporting more synchronized and efficient team behavior. Concerning the second sub-research question, studies from adjacent domains show that teams undergoing TMM-focused preparation generate mental models that align more closely with expert expectations, particularly in simulated or high-fidelity task environments. These findings, though not originating directly from incident response settings, provide a plausible theoretical and practical basis for applying TMM training to cybersecurity contexts. Thus, the literature collectively supports the relevance and potential utility of TMM-specific training for improving the effectiveness of incident response teams, while underscoring the need for further empirical investigation tailored to cybersecurity operations.

Drawing upon these insights, this study investigates whether targeted TMM training can enhance the effectiveness of *ad hoc* cybersecurity incident response teams. Specifically, the **Main Research Question (MRQ)** asks whether a structured TMM intervention improves overall team effectiveness in these time-constrained, high-stakes environments. To operationalize this inquiry, two sub-research questions are posed: **Sub-Research Question 1 (SRQ1)** asks whether the intervention increases the similarity of team members' mental models compared to conventional cybersecurity training, while **Sub-Research Question 2 (SRQ2)** examines whether it improves the accuracy of those models, aligning them more closely with expert expectations. These questions, grounded in the literature reviewed, guide the methodological design described in the following section.

4. Methodology

4.1 Research Design and Exercise Setup

This study employed a quasi-experimental pre-test/post-test control group design to investigate the impact of a targeted training intervention on the development of Team Mental Models (TMM) within *ad hoc* cybersecurity incident response (IR) teams. The core hypothesis posits that enhancing TMM similarity and accuracy through training can improve IR team processes and performance, drawing upon established links between shared mental models and team effectiveness [21], [24], [46].

The experimental setup involved two distinct teams, designated Team A (Control Group) and Team B (Experimental Group), participating in a simulated IR tabletop exercise. The use of tabletop exercises provides a controlled yet realistic environment to observe team dynamics and decision-making processes in response to simulated cybersecurity incidents, a common approach in IR training and research [77].

This pre-post design with a control group allows for comparison of changes within each team (pre vs. post) and between the teams (experimental vs. control), helping to isolate the specific effect of the TMM-focused training intervention [34]. The use of two different scenarios (A and B) aimed to mitigate learning effects specific to a single scenario, although potential differences in scenario difficulty remain a consideration.

Each team consisted of 9 people, and one of them were the team leader as incident commander. Team leaders are not elected but assigned by the seniority -department managers. The exercise is executed in English language. The team composition is not balanced but randomized for the quasi-experiment, however the small group size has limitations on balancing properly to be explained in Discussion sections.

In order to let the participants focus on the exercise, the exercise held in a location out of office building but close to it. No laptops were allowed to prevent distraction, except for the presentation laptop facilitator used to present. Each team member signed the consent form for individual consent, provided both in English and Estonian language (see Appendix 7 and 7). A separate form is signed by a company official to allow usage of this exercise as an academic material (see Appendix 7). Control group started at 09.30 and finished around 12.30, while experimental group started at 14.00 and finished round 17.00. The

team leaders were given the incident response playbooks, and related policies printed out.

The exercise followed a structured sequence:

1. **Pre-Training Scenario (Scenario A):** Both teams independently participated in the first tabletop scenario, simulating a specific cybersecurity incident. This phase established a baseline measure of TMM similarity and accuracy for each team before any intervention.
2. **Intervention Phase:** Following Scenario A, a break was provided. During this break:
   - **Team A (Control Group):** Received a pseudo-training session focused on general security hygiene principles, unrelated to TMM development. This controlled for potential Hawthorne effects or the simple effect of receiving any training.
   - **Team B (Experimental Group):** Received a targeted training intervention specifically designed to enhance TMM development. This training provided explicit methods and strategies for teams to build shared understanding regarding task requirements, team roles, situational awareness, and communication protocols, consistent with principles of team training aimed at improving cognitive states [59], [84], [85].

     As part of the training, participants were introduced to strategies outlined in Section 3.4 *Strategies for Developing Team Mental Models* to support team processes during the incident. While the tabletop exercise provided a simulation-based environment, additional support was offered through structured elements such as strategy briefings, debriefings, and the use of storytelling as practical tools for enhancing shared understanding. Given that the team operated in a co-located setting, communication protocols and physical tools—such as whiteboards, post-it notes, and related materials—were incorporated into the simulation setup. These resources also served as integral components of the storytelling approach, helping to externalize and spatially represent the evolving team mental model throughout the scenario.

     Due to the randomized composition of the teams, responsibilities had to be redistributed at the outset of the exercise, with the exception of the team leader, who was assigned based on staff seniority. In this context, cross-training did not function as an optional team development strategy but rather became a necessary task, particularly in cases where the team leader needed to reassign key roles using available personnel.
3. **Post-Training Scenario (Scenario B):** Both teams independently participated in a second, distinct tabletop scenario. This phase measured TMM similarity and

accuracy after the intervention, allowing for assessment of training impact.

4.1.1 Scenarios

Scenario A The scenario A is a data breach scenario based on an initial vector provided by a human error. In order to provide a simulated environment, the attack is tested on a lab environment and actual logs are generated. Then, the user name, host name, IP addresses and similar details are replaced with the scenario-related information. Logs are provided in a single HTML-page (see Figure 5) that imitates a SIEM interface.

The attacker manages to exfiltrate cardholder data (CHD) from the company. "At a minimum, cardholder data consists of the full PAN. Cardholder data may also appear in the form of the full PAN plus any of the following: cardholder name, expiration date and/or service code" [86], while PAN is an "[a]cronym for “primary account number.” Unique payment card number (credit, debit, or prepaid cards, etc.) that identifies the issuer and the cardholder account".

The incident begins with a compromised VS Code (an integrated development environment (IDE)) extension update, leading to the installation of a malicious service that establishes a persistent tunnel using VS Code’s Remote Tunnels feature. The attacker performs initial reconnaissance; current user, environment variables, computer information, network information, and checks for known folders for possible sensitive data. Attacker then discovers basic information in the developer’s source code and configurations about names and addresses of servers in the network. Afterwards, attacker finds credentials in a Notepad++ (an open source text editor) backup file, uses these credentials for accessing to a production server using file shares user has access to, downloads and compresses the files, and finally exfiltrates cardholder data files through the established tunnel. Key indicators include Windows event logs on privileged service call and process creation (`Security-4673` and `Security-4688`, `Sysmon-1`), service installation (`System-7045`), specific process/command line activity (`Security-4688`, `Sysmon-1`), PowerShell Script-Block logs (`PowerShell-4104`), and firewall logs showing large POST requests to the tunnel domain.

The expected response involves immediate isolation of the compromised developer machine and the server, triaging the incident by reviewing logs (EDR, firewall, SIEM), identifying and removing the persistence mechanism (malicious service), scoping the incident (data accessed/exfiltrated), revoking compromised credentials, preserving forensic evidence (images, logs), identifying the root cause (VS Code extension), reporting, remediation (hardening, monitoring), and triggering compliance processes (PCI DSS due to cardholder

Discover

Search

Refresh

Field	Value
t agent.id	199
t agent.name	LAP99
t win.eventdata.commandLine	"C:\Program Files\Microsoft VS Code\bin\..\Code.exe" "C:\Program Files\Microsoft VS Code\bin\..\resources\app\out\cli.js" tunnel service install --accept-server-license-terms
t win.eventdata.company	Microsoft Corporation
t win.eventdata.currentDirectory	C:\Users\johndoe\Downloads\
t win.eventdata.description	Visual Studio Code
t win.eventdata.fileVersion	1.99.3
t win.eventdata.hashes	MD5=F27FD16B5B9CA9DE05FE87C38097F323,SHA256=485F2A72CBE44B9ECF3C35E6F59DF5DDC13861EF54E33E19E4F1E55E76F23224,IMPHASH=41194EA2F3E991657107EBDBF563F034
t win.eventdata.image	C:\Program Files\Microsoft VS Code\Code.exe
t win.eventdata.integrityLevel	Medium
t win.eventdata.logonGuid	{480c9770-1e9e-6805-ed05-1c0000000000}
t win.eventdata.logonId	0x1c05ed
t win.eventdata.originalFileName	electron.exe
t win.eventdata.parentCommandLine	C:\WINDOWS\system32\cmd.exe /c ""C:\Program Files\Microsoft VS Code\bin\code.cmd" tunnel service install --accept-server-license-terms"
t win.eventdata.parentImage	C:\Windows\System32\cmd.exe
t win.eventdata.parentProcessGuid	{480c9770-21e2-6805-5a03-00000000e402}
t win.eventdata.parentProcessId	17704
t win.eventdata.parentUser	NETSEE\johndoe
t win.eventdata.processGuid	{480c9770-21e2-6805-5b03-00000000e402}
t win.eventdata.processId	12908
t win.eventdata.product	Visual Studio Code
t win.eventdata.terminalSessionId	1
t win.eventdata.user	NETSEE\johndoe
t win.eventdata.utcTime	2025-04-20 16:33:38.499
t win.system.channel	Microsoft-Windows-Sysmon/Operational
t win.system.computer	LAP99
t win.system.eventID	1
t win.system.eventRecordID	12154103
t win.system.keywords	0x8000000000000000
t win.system.level	4
t win.system.message	Process Create:\nRuleName: - UtcTime: 2025-04-20 16:33:38.499 ProcessGuid: {480c9770-21e2-6805-5b03-00000000e402} ProcessId: 12908 Image: C:\Program Files\Microsoft VS Code\Code.exe FileVersion: 1.99.3 Description: Visual Studio Code Product: Visual Studio Code Company: Microsoft Corporation OriginalFileName: electron.exe CommandLine: "C:\Program Files\Microsoft VS Code\bin\..\Code.exe" "C:\Program Files\Microsoft VS Code\bin\..\resources\app\out\cli.js" tunnel service install --accept-server-license-terms CurrentDirectory: C:\Users\johndoe\Downloads\ User: NETSEE\johndoe LogonGuid: {480c9770-1e9e-6805-ed05-1c0000000000} LogonId: 0x1C05ED TerminalSessionId: 1 IntegrityLevel: Medium Hashes: MD5=F27FD16B5B9CA9DE05FEB7C38097F323,SHA256=485F2A72CBE44B9ECF3C35E6F59DF5DDC13861EF54E33E19E4F1E55E76F23224,IMPHASH=41194EA2F3E991657107EBDBF563F034 ParentProcessGuid: {480c9770-21e2-6805-5a03-00000000e402} ParentProcessId: 17704 ParentImage: C:\Windows\System32\cmd.exe ParentCommandLine: C:\WINDOWS\system32\cmd.exe /c ""C:\Program Files\Microsoft VS Code\bin\code.cmd" tunnel service install --accept-server-license-terms" ParentUser: NETSEE\johndoe
t win.system.opcode	0
t win.system.processID	6408
t win.system.providerGuid	{5770385f-c22a-43e0-bf4c-06f5698ffbd9}
t win.system.providerName	Microsoft-Windows-Sysmon
t win.system.severityValue	INFORMATION
t win.system.systemTime	2025-04-18T16:00:38.5023443Z
t win.system.task	1
t win.system.threadID	7500
t win.system.version	5
t @timestamp	2025-04-18T16:00:38.5023443Z

Field	Value
t agent.id	199
t agent.name	LAP99
t win.eventdata.accountName	LocalSystem
t win.eventdata.imagePath	C:\Program Files\Microsoft VS Code\bin\code-tunnel.exe
t win.eventdata.serviceName	VS Code Remote Tunnel service
t win.eventdata.serviceType	user mode service
t win.eventdata.startType	auto start
t win.system.channel	System
t win.system.computer	LAP99
t win.system.eventID	7045
t win.system.eventRecordID	406117
t win.system.keywords	0x8080000000000000
t win.system.level	4
t win.system.message	A service was installed in the system. Service Name: VS Code Remote Tunnel service Service File Name: "C:\Program Files\Microsoft VS Code\bin\code-tunnel.exe Service Type: user mode service Service Start Type: auto start Service Account: LocalSystem
t win.system.opcode	0
t win.system.processID	556
t win.system.providerGuid	{555908d1-a6d7-4695-8e1e-26931d2012f4}
t win.system.providerName	Service Control Manager
t win.system.severityValue	INFORMATION
t win.system.systemTime	2025-04-18T16:02:19.2127764Z
t win.system.task	0
t win.system.threadID	19324
t win.system.version	0
t @timestamp	2025-04-18T16:02:19.2127764Z

Figure 5. A screenshot of the simulated SIEM dashboard showing remote tunnel installation log

data).

Scenario B Scenario B is based on a real-life incident which has a great writeup written by a security company [87]. It consists of a ransomware attack chaining several vulnerabilities on VMware and ESX platform. For this scenario, no logs are provided as SIEM server is also encrypted, making it significantly harder than Scenario A. The team had to work together to create a timeline of incidents and mental model with no visual helpers.

The incident starts with a web shell (`shell.php`) uploaded to a web VM, detected by file integrity monitoring. The attacker then exploits a vulnerability (CVE 2025-22224) to escape the VM and compromise the underlying hypervisor. From the hypervisor, the attacker deploys and executes ransomware, encrypting multiple VMs (including critical ones like the SIEM) residing on a shared datastore. The attack culminates in a ransom note appearing in vCenter. Key indicators include firewall logs (POST to `upload.php`), file integrity alerts, ESX host logs (`vmkernel.log`, process activity), increased storage activity, service failures on encrypted VMs, datastore logs showing file renames, and the ransom note.

The expected response prioritizes halting the ransomware spread *without shutting down* the compromised hypervisor initially to allow for live forensics. This involves isolating ESX host network interfaces and unmounting the datastore. Live forensics (memory dump, processes, network connections) should be performed *before* acquiring disk images. Triage involves confirming the web shell, ESX host compromise, and scope of encryption. Since the SIEM is likely compromised, alternative logging sources must be used. Recovery involves restoring from backups. Communication, reporting, remediation (patching/rebuilding ESX host, hardening), and secondary containment of the initial web VM are also crucial.

4.2 Measurement of Team Mental Models (TMM)

Measuring the complex construct of TMM requires a multi-faceted approach, as recommended by measurement literature [88], [27]. This study employed distinct methods to capture both the **similarity** (convergence of understanding among team members) and **accuracy** (correspondence of team understanding with expert knowledge) of TMMs, leveraging the data collected during the pre- and post-training tabletop exercises.

4.2.1 Measuring TMM Similarity

The measure of team mental model similarity is measured by questionnaire items was evaluated by analyzing each Likert-scale response using parametric tests appropriate for repeated measures. Repeated-measures ANOVA was applied to assess the main effects of time (pre- vs. post-training), group (control vs. experimental), and their interaction, reporting F-statistics, p-values, and ω^2 effect sizes. Where the omnibus ANOVA indicated directional trends or significant effects, follow-up analyses were conducted using paired or independent-samples t-tests. Bonferroni correction was applied where multiple comparisons were required, and Cohen's d with 95% confidence intervals was reported. This approach aligns with established practices in TMM measurement using perceptual similarity metrics, as outlined in [22], [57] and discussed in Section 3.5.

For the free-text incident-report narratives, this study adopted an embedding-based cohesion measure derived from Sentence-BERT (SBERT). Each participant's full response was encoded into a single 384-dimensional vector and compared to all peers via cosine similarity; the resulting row-wise average yields one "cohesion" score per person per timepoint. Within-subject changes in cohesion for the experimental group were tested with the Wilcoxon signed-rank test (rank-biserial effect size), and between-group differences in change scores ($\Delta = \text{post} - \text{pre}$) were evaluated using the Mann–Whitney U test (rank-biserial effect size). Embedding-based semantic comparison offers a structural method to complement perceptual metrics, addressing the trade-off between subjective and objective techniques noted in [22], [71].

By combining repeated-measures ANOVA and targeted t-tests for structured questionnaire data with SBERT-based cohesion scoring and nonparametric hypothesis tests for free-text data, this methodology provides a robust, assumption-aware assessment of Team Mental Model similarity and its evolution under different training conditions. This mixed-method triangulation directly responds to the call in TMM literature for multimodal approaches to mitigate mono-method bias and to capture both explicit and latent model structures [22], [27], [57].

4.2.2 Measuring TMM Accuracy

Sub-Research Question 2 (SRQ2) focuses on whether Team Mental Model (TMM) training improves the accuracy of teams' shared understanding of incident scenarios. In our design, each team appointed a department head as its leader based on seniority—rather than election—who was responsible for producing a single, team-level incident report for each

scenario. All other participants completed an identical report template individually, but team leaders were explicitly instructed to consult with their members to craft a more accurate, collective representation.

TMM accuracy was measured as the degree of correspondence between each team's overall report and an objectively defined benchmark model. Drawing on Mathieu *et al.* [24] and Cannon-Bowers *et al.* [21], accuracy reflects how closely a team's shared mental model matches an expert or "Gold Standard" model of the task environment. Rather than relying on holistic expert ratings—which may conflate conceptual understanding with writing style—this study operationalizes accuracy through structured content analysis of the complete team report. This follows the guidance of [57] and addresses concerns regarding construct validity raised in Section 3.5.

For each scenario (A and B), a scenario-specific Gold Standard Model was first synthesized by consolidating all critical facts, decision points, and prescribed actions from the scenario description and response plan. Based on this benchmark, a ten-item checklist was developed to target verifiable elements in the team reports—such as identification of the initial attack vector, recognition of hypervisor compromise, and recommendation of system isolation. Each checklist item was designed to be scored as present or absent, eliminating the need for subjective interpretation of prose quality. This structured approach responds to the critique that many TMM accuracy assessments suffer from unclear or subjective scoring frameworks [27], [67].

This approach was chosen because it directly operationalizes TMM accuracy against an objective, scenario-relevant standard and focuses on concrete report content rather than writing fluency. The structured checklist enhances inter-rater consistency, minimizes bias due to report completeness, and remains feasible within the scope of a master's thesis. Application of these checklists to the four team reports produced quantitative accuracy scores, which were analyzed using a Difference-in-Differences (DiD) framework. This approach compared pre-to-post changes in the experimental group against those in the control group, enabling the isolation of the effect of TMM training on teams' mental model accuracy. In doing so, this measurement method addresses common challenges in TMM research, including issues of reliability, ecological validity, and scoring fidelity discussed in Section 3.5 and in prior critiques [22], [57], [73].

5. Results

The first subsection employs t-tests and ANOVA to compare within-group and inter-group result similarities to assess Sub-Research Question (SRQ) 1. It is tested if the team intervention via team training improved TMM similarity within the experimental group. The similarity results are going to be compared between control group and experimental group for pre and post-training setups.

The second section addresses SRQ 2, examining whether the TMM-specific training intervention led to a greater improvement in Team Mental Model (TMM) accuracy compared to the general security training. TMM accuracy was measured using the custom checklists derived from the Gold Standard Models for each scenario, applied to the team leader reports. Scores represent the number of key elements correctly identified or addressed out of a possible 10 for each scenario.

5.1 Similarity

In similarity assessments, two distinct approaches were employed corresponding to two different types of questions. First, Likert-scale items were analyzed using JASP[89] to perform t-tests and ANOVA. Second, free-text responses in the reports were examined for semantic similarity. For this analysis, Python code and Sentence-BERT (SBERT) were utilized.

5.1.1 Likert scale questions

t-tests

The paired samples t-tests, given in Table 4, revealed diverse patterns of change in participant scores across the training period. Notably, Questions 1d, 1e, 2a, and 2b showed statistically significant declines post-training, with large to very large effect sizes (Cohen's d = 0.95 to 2.28), suggesting meaningful performance deterioration on those items. These findings warrant careful attention, as they may reflect unintended consequences of the intervention, measurement artifacts, or topic-specific challenges.

Encouragingly, several items showed positive trends, even if not statistically significant. Question 1b, for instance, showed a moderate improvement ($t(16) = 1.692$, $p = 0.110$,

$d = -0.827$), and 1a and 6b showed small non-significant gains, with effects in the expected direction. While these did not reach the conventional $p < 0.05$ threshold, their effect sizes and direction suggest potential for improvement with a larger sample or adjusted training content.

Table 4. Paired Samples T-Test Results Across All Questions

Question	t	df	p	Cohen's d	95% CI (d)	Interpretation
1a	0.212	15	0.835	-0.438	[-1.369, 0.494]	Small increase post-training; not statistically significant.
1b	1.692	16	0.110	-0.827	[-1.861, 0.208]	Moderate improvement; $p > 0.05$, CI includes 0.
1c	-0.270	17	0.790	0.127	[-0.784, 1.038]	Small decrease; not significant.
1d	-3.557	17	0.002	1.498	[0.559, 2.437]	Significant decline post-training. Large effect.
1e	-3.439	17	0.003	1.448	[0.519, 2.376]	Significant decline. Large effect.
2a	-5.477	15	$< .001$	2.280	[1.191, 3.369]	Strong significant decline. Very large effect.
2b	-2.266	16	0.038	0.945	[0.051, 1.838]	Statistically significant decline. Large effect.
2c	-1.683	16	0.112	0.702	[-0.180, 1.585]	Moderate decline; not statistically significant.
6a	-1.417	16	0.176	0.591	[-0.290, 1.472]	Moderate decline; not statistically significant.
6b	0.436	16	0.668	-0.182	[-1.063, 0.698]	Small non-significant improvement.

Presented in Table 5, independent samples t-tests compared post-training performance between control and experimental groups. None of the comparisons reached statistical significance, but several comparisons showed moderate to large effect sizes, particularly for Questions 1a ($d = 0.776$) and 1b ($d = -0.956$). The wide confidence intervals for these effects included zero, indicating that although the magnitude of the observed differences is notable, statistical certainty is lacking—potentially due to limited sample size or high variance.

Table 5. Independent Samples T-Test Results (Post-Test Comparison Between Groups)

Question	t	df	p	Cohen's d	95% CI (d)	Interpretation
1a	-1.602	14	0.132	0.776	[-0.266, 1.818]	Moderate group difference favoring experimental; not significant.
1b	2.012	15	0.063	-0.956	[-2.016, 0.104]	Large effect favoring control; close to significant.
2a	1.633	16	0.122	-0.756	[-1.722, 0.211]	Moderate effect favoring control; not significant.
2b	-0.351	16	0.730	0.165	[-0.725, 1.054]	Minimal difference; not significant.

Overall, these results highlight both positive indicators—like upward trends in some questions—and areas of concern, particularly where training appears to coincide with a significant performance drop. Importantly, the presence of moderate effects with non-significant p-values emphasizes the value of effect size interpretation and the need for continued monitoring or refinement of the training program.

Repeated Measures ANOVA Tests

To assess the impact of the training intervention on questionnaire responses, repeated measures ANOVA was applied independently to each item. This approach allowed for the evaluation of three distinct effects: the main effect of time (pre-training vs. post-training), the main effect of group (control vs. experimental), and the interaction between time and group. The use of repeated measures improves statistical power by accounting for within-subject variance [90]. For each item, within-subjects results report F-statistics, p-values, and effect sizes using omega squared (ω^2), which provides a less biased estimate of explained variance in small samples [91]. Between-subjects effects are also included to examine overall differences between groups irrespective of time.

To aid interpretation, descriptive statistics are presented for each group at both time points, including means, standard deviations, and standard errors. In cases where directional trends were observed, line plots are provided to illustrate temporal patterns. Additionally, post hoc comparisons between groups at the post-test stage were conducted using independent-samples t-tests. These tests report mean differences, t-values, Bonferroni-adjusted p-values to correct for multiple comparisons, and effect sizes using Cohen's d, along with 95% confidence intervals for the magnitude of effects. Effect size reporting and confidence intervals are essential for assessing the practical relevance of findings beyond p-value thresholds [91], [92]. This combination of inferential testing, effect quantification, and visualization facilitates a robust and interpretable analysis of item-level training effects.

To assess the effects of the training intervention across multiple questionnaire items, repeated-measures ANOVA was employed as the primary inferential framework. This allowed us to evaluate both the main effect of time (pre- vs. post-training) and the interaction between time and group (control vs. experimental). For each question, it was reported whether the change over time reached statistical significance and whether any group-specific patterns were observed. The ANOVA results were complemented by descriptive statistics to identify trends in group means, and post hoc independent samples t-tests were applied to compare group scores at post-test. These post hoc tests allowed for clearer interpretation of group-level differences, even when omnibus effects were not significant, particularly through effect size estimates and directional changes.

Comprehensive, question-by-question ANOVA tables—including F-statistics with degrees of freedom, exact p-values, and ω^2 effect sizes—alongside full post hoc t-test outputs (mean differences, Bonferroni-adjusted p-values, 95% confidence intervals, and Cohen's d) and corresponding descriptive statistics and line plots are provided in Appendix 7. This material offers the granular evidence underpinning the summary results in Table 6.

Table 6 presents a comprehensive overview of these findings. Each row corresponds to a single question and reports the p-values for time and interaction effects, the general trend in score direction, the interpretation of the post hoc comparison, and a final synthesized assessment. To enhance interpretability, a categorical classification scheme was introduced to distinguish between significant positive or negative effects, near-significant changes, and directional trends lacking statistical significance. This allows for greater resolution in evaluating practical training outcomes. For instance, while Question 1d showed no group interaction, it revealed a strong and statistically significant drop in scores overall—confirmed by descriptive statistics and post hoc testing. Conversely, Question 1b exhibited a promising upward trend in the experimental group, with near-significance in the interaction term and a moderate effect size, indicating a potential training benefit that may warrant further investigation in future studies.

Table 6. Summary of Repeated Measures ANOVA, Post Hoc Tests, and Final Assessment

Q	Time Effect (p)	Time × Team (p)	Direction	Post Hoc Interpretation	Final Assessment	Result Classification
1a	0.681	0.133	Mixed	Moderate effect (not significant)	No significant change; slight shift in direction between groups	No Change
1b	0.179	0.055	Positive (Exp)	Moderate effect (not significant)	Positive trend in experimental group; interaction nearly significant	Near-Significant Positive Change
1c	0.778	0.743	Positive (Exp)	Not significant	Mild experimental gain; not statistically meaningful	Positive Trend (Not Significant)
1d	0.001	0.549	Negative	Not significant	Significant decline across both groups; moderate effect confirmed	Significant Negative Change
1e	0.001	0.949	Negative	Not significant	Significant drop; consistent with post hoc and descriptives	Significant Negative Change
2a	<.001	0.471	Negative	Moderate effect (not significant)	Strong decline observed; consistent across all metrics	Significant Negative Change
2b	0.039	0.543	Negative	Small effect; not significant	Moderate drop; confirmed by both ANOVA and post hoc	Significant Negative Change
2c	0.099	0.834	Mixed	Small trend; not significant	No significance, but opposing trends in descriptives	No Change
6a	0.178	0.672	Negative	Small effect; not significant	Minor decline; no significant effect across metrics	Non-Significant Negative Trend
6b	0.641	0.538	Neutral	Small effect; not significant	Scores stable; no meaningful change detected	No Change
6c	0.285	0.422	Negative (Exp)	Small effect (Exp dropped slightly)	Slight drop in experimental group; no statistical significance	Non-Significant Negative Trend
6d	0.608	0.844	Neutral	Minimal difference; stable scores	Fully stable results across all metrics; no change	No Change

5.1.2 Free-text Similarity

Participants described incident timelines in their own words using a free-text section, allowing us to capture their mental models. For the assessment of free-text similarity, Sentence-BERT (SBERT) was selected as the underlying embedding model[93].

SBERT is a Siamese extension of the Transformer-based Bidirectional Encoder Representations from Transformers (BERT) architecture, developed to produce semantically meaningful, fixed-size embeddings for entire sentences or short texts. Unlike models that operate at the token level, SBERT captures sentence-level semantics, making it particularly useful for handling variations in phrasing, domain-specific terminology, and paraphrasing—common features of incident-report narratives. This capability ensures that semantically similar statements are mapped to nearby vectors in the embedding space. SBERT operates by using two identical BERT encoders with shared weights to independently process pairs of input sentences. The resulting token-level outputs are then aggregated into a single vector, typically via mean pooling over the final hidden states. During fine-tuning, SBERT is trained on tasks such as natural language inference or semantic textual similarity, using classification or regression objectives to align cosine similarity scores with human-annotated semantic judgments. This training process enables SBERT to represent subtle differences in meaning and contextual similarity, making it well-suited for measuring cohesion in textual data.

Free-text incident reports capture nuanced descriptions of events, often featuring inconsistent terminology and varied phrasing; comparing their semantic content therefore requires embeddings that balance high accuracy with computational efficiency. By producing sentence embeddings that accurately reflect semantic content, SBERT allows direct comparison of report pairs through cosine-similarity measures, supporting tasks such as clustering, retrieval, and statistical analysis.

The *all-MiniLM-L6-v2* variant of SBERT was selected to balance semantic precision with computational efficiency [94]. This particular architecture was not chosen based on corpus size, but rather for its demonstrated capability to produce high-quality sentence embeddings that effectively capture subtle semantic relationships, even within small datasets. Previous studies have shown that SBERT consistently outperforms standard BERT models on semantic textual similarity benchmarks, requiring minimal fine-tuning [93]. These characteristics make SBERT particularly suitable for evaluating cohesion in free-text incident reports, where the objective is to assess conceptual consistency rather than to train a language model from the ground up.

This model produces 384-dimensional embeddings while occupying only 22 MB on disk, in contrast to BERT-base's 420 MB footprint. Despite its compact size, MiniLM retains over 99 % of BERT-base's performance on extractive QA (SQuAD 2.0) and the GLUE suite, using only half of the Transformer parameters and FLOPs, and yields a 2.7× inference speed-up on GPU hardware (e.g. processing upwards of 2 000 sentences per second). These characteristics make *all-MiniLM-L6-v2* particularly well-suited for large-scale, interactive analyses of free-text incident reports.

Tests Each participant's free-text response was reduced to a single SBERT-based cohesion score by encoding the complete response into a 384-dimensional embedding vector e_i, followed by computing the average cosine similarity between e_i and all other participants' embeddings within the same experimental condition. Formally, for N valid responses in a condition,

$$\mathrm{score}_i \;=\; \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} \cos\bigl(e_i, e_j\bigr).$$

This procedure produces one independent observation per person, thereby satisfying the assumptions of paired and independent nonparametric tests.

For each participant, the pre-training cohesion score pre_i and the post-training score post_i were recorded, and the change score was computed as:

$$\Delta_i \;=\; \mathrm{post}_i \;-\; \mathrm{pre}_i.$$

To assess whether Team Mental Model (TMM) training induced a within-subject improvement, a paired Wilcoxon signed-rank test was applied to compare pre_i and post_i scores within the experimental group. To isolate the training effect from potential practice effects or scenario familiarity, the distributions of Δ_i were further compared between control and experimental participants using an independent-samples Mann–Whitney U test.

This participant-level framework provides a transparent and assumption-light evaluation of semantic convergence. By reducing each report to a single cohesion score and testing pre-post and between-group differences using nonparametric methods, the analysis directly quantifies whether TMM training produced a statistically reliable increase in participants' alignment of incident-report language.

SBERT calculations were implemented in Python using the *SentenceTransformers* library. The complete source code is available in Appendix 7. Then, the extracted and computed

fields -namely *participant_id, group*, *pre_score*, *post_score*, and *delta_score*- are exported into a CSV file for further analysis in JASP.

Table 7 reports the participant-level SBERT cohesion scores (mean cosine similarity to all peers) before and after training for both control and experimental groups. Figure 6 shows the group-level means before and after training, with error bars omitted for clarity.

Table 7. Participant-Level SBERT Cohesion Scores by Group and Timepoint

Group	**Pre-training**	**Post-training**
Control	0.4764	0.4480
Experimental	0.4633	0.4785

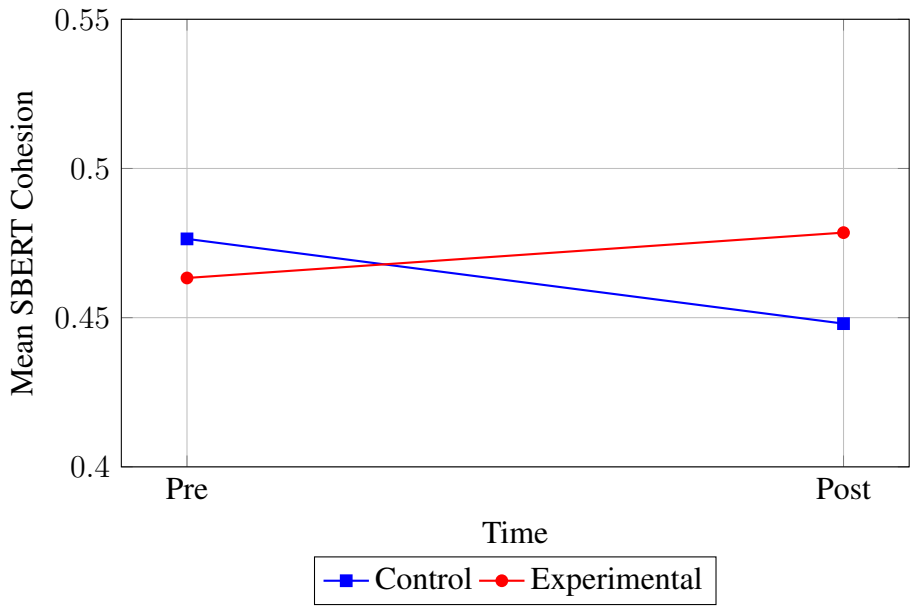

Figure 6. Mean SBERT cohesion scores before and after training for control and experimental groups.

Pre- vs. post-training cohesion within each group was tested using the Wilcoxon signed-rank test. Results are shown in Table 8.

Table 8. Wilcoxon Signed-Rank Tests of Pre vs Post Cohesion

Comparison	**W**	**p**	**Rank-biserial** r
Control (pre vs post)	45.0	0.423	–0.250
Experimental (pre vs post)	18.0	0.456	0.265

Neither the control refresher nor the Team Mental Model training produced a statistically

significant within-participant change in cohesion ($p > 0.4$ in both cases).

Each participant's change score (Δ = post − pre) was subsequently computed and compared between groups using the Mann–Whitney U test. Descriptive means appear in Table 9, and the test results in Table 10.

Table 9. Descriptive Statistics of Change Scores Δ

Group	**Mean Δ**	**Median Δ**
Control	-0.0284	-0.0027
Experimental	$+0.0152$	-0.0254

Table 10. Mann–Whitney U Test on Change Scores Δ

Test	**U**	**p**	**Rank-biserial r**
Experimental vs Control Δ	20.0	0.397	0.286

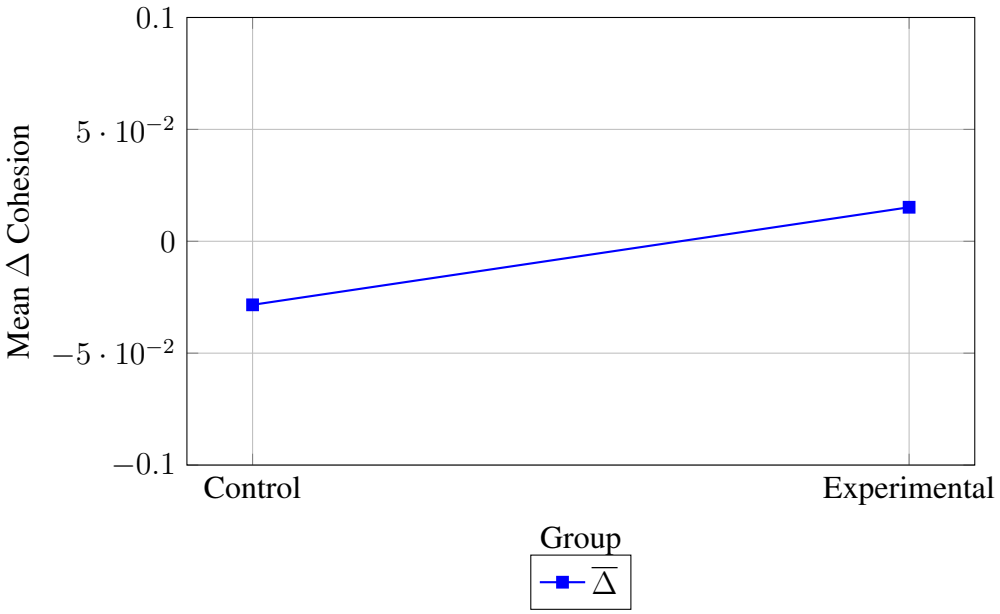

Figure 7. Mean change in SBERT cohesion (Δ) by group.

Figure 7 plots the group means of Δ; control shows a slight decrease, experimental a slight increase, but the difference is non-significant.

Although the experimental group's mean cohesion rose descriptively from 0.4633 to 0.4785, the Wilcoxon signed-rank test was non-significant (W = 18.0, p = 0.456, r = 0.265). The control group's mean fell slightly, also non-significant (W = 45.0, p = 0.423, r =

-0.250). Comparing change scores, the experimental cohort's modest gain (+0.0152) did not significantly exceed the control's loss (-0.0284) (U = 20.0, p = 0.397, r = 0.286). These results indicate that, with the current sample, Team Mental Model training did not produce a reliably greater improvement in semantic cohesion than the control intervention, suggesting that observed fluctuations may stem from practice effects or individual variability rather than the TMM curriculum.

5.2 Accuracy

This section addresses Sub-Research Question 2 (SRQ2) by examining whether the team mental model (TMM)–specific training produced a greater improvement in TMM accuracy than the general security training. Accuracy was operationalized through custom checklists derived from the Gold-Standard Models for each scenario; the resulting scores, which range from 0 to 10, reflect the number of key elements correctly identified in the team-leader reports.

5.2.1 Accuracy Scores and Within-Team Changes

Table 11 reports the accuracy scores for both teams across the two scenarios together with the within-team change from Scenario A (pre-training) to Scenario B (post-training). Because the checklist is capped at ten points, each one-point change represents a 10 percent shift in the proportion of Gold-Standard elements captured.

Table 11. TMM accuracy scores (0–10 scale) and corresponding within-team changes

Team	**Scenario A (Pre)**	**Scenario B (Post)**	**Change (Δ)**
Control	5.0	4.5	-0.5
Experimental	4.5	5.5	$+1.0$

The baseline scores in Scenario A were comparable, with control group recording 5.0 points and experimental group 4.5 points. After training, control group's accuracy declined to 4.5, yielding a within-team change of -0.5 points (a 5 percent decrease), whereas experimental group's accuracy increased to 5.5, corresponding to a within-team gain of $+1.0$ points (a 10 percent increase). Although both teams still captured only about half of the Gold-Standard content, experimental group exhibited a clear post-training improvement while control group did not.

5.2.2 Difference-in-Differences Analysis

To isolate the relative effect of the TMM-specific training, a difference-in-differences (DiD) estimate was computed by subtracting the change observed in the control group from that observed in the experimental group:

$$DiD = \Delta_{\text{experimental group}} - \Delta_{\text{control group}} = (+1.0) - (-0.5) = +1.5. \tag{5.1}$$

The positive DiD value of $+1.5$ indicates that, relative to the control group, the TMM-specific training was associated with an additional 1.5-point improvement in checklist accuracy. On the 0–10 scale, this corresponds to an incremental gain of 15 percent of the total possible score. Because the checklist length is constant across teams and scenarios, any linear rescaling (for example, expanding the checklist to 20 items and doubling all raw scores) would leave the sign and proportional meaning of the DiD unchanged.

6. Discussion

6.1 Study Aim and Theoretical Framing

This study presents a novel investigation across three complementary domains: Team Mental Models (TMMs), task teams, and cybersecurity incident response (IR). Although TMM theory has been extensively examined in military, aviation, and healthcare settings [21], [24], its application to the unique socio-technical environment of cybersecurity IR constitutes a significant theoretical extension. The cybersecurity domain—with its highly technical and rapidly evolving threats, abstract risk landscape, and need for diverse expertise—pushes established TMM principles to their limits, demanding rapid, shared understanding under extreme complexity, ambiguity, and time pressure.

Temporary task teams and *ad hoc* IR cells alike struggle to develop shared cognition when members have limited prior collaboration and only minutes to align on procedures [14], [17], [20]. While structured interventions can mitigate these barriers [35], most prior work examines programs lasting hours or longer—leaving open the question of how quickly TMMs can form in genuine emergency contexts. We therefore ask: *How can shared mental models be accelerated to operational effectiveness in teams assembled reactively to emerging cyber threats?*

Our Main Research Question (MRQ) investigates whether a structured TMM protocol—comprising cross-training and strategy briefing—can enhance overall team effectiveness in *ad hoc* IR teams. To operationalize this, we contrasted two nine-member teams (intervention vs. control) in a quasi-experimental exercise. Measurement combined Likert-scale similarity ratings and SBERT-based semantic cohesion of free-text summaries with expert-derived checklists aligned to Endsley's situational-awareness levels [30], [93]. Although neither similarity nor accuracy improvements reached statistical significance, the observed descriptive shifts delineate the boundary conditions of micro-intervention efficacy. Beyond theoretical contributions, this research offers a replicable mixed-method pipeline—integrating semantic embeddings, expert checklists, and surveys—that can be adopted for future studies of rapid team-cognition development in high-stakes, time-sensitive settings.

Controlled quasi-experiments with *ad hoc* incident-response teams remain rare within cybersecurity research. Our study therefore offers incremental theoretical and methodological

insights while avoiding overstatement of its scope. On the theoretical side, we test whether shared-cognition constructs, developed for stable organizations, remain relevant in volatile incidents, and we challenge assumptions about the time needed for model convergence. Methodologically, we integrate SBERT-based semantic cohesion with expert-benchmarked checklists and conventional surveys in a mixed-method pipeline that balances automation with rigorous validation. This design aims to capture both implicit and explicit facets of team cognition while providing a scalable template for future work where rapid alignment is mission-critical.

The research questions translate these design choices into testable claims about similarity and accuracy. SRQ1 investigates whether embedding metrics detect meaningful convergence of mental models immediately after a brief intervention in a live incident-response exercise. SRQ2 evaluates whether those converged representations align with standards produced by experienced responders, thereby gauging the practical value of concise team-mental-model protocols under pressure. Together, the findings offer a cautious lower-bound estimate of micro-intervention effectiveness in cybersecurity incidents and refine theory on how shared models stabilize when decisions unfold within minutes.

6.2 Effects on Team Mental Model similarity (SRQ1)

Analysis of the nine-item Likert instrument and of free-text SBERT cohesion scores yielded no statistically significant interaction effects attributable to the TMM briefing. This finding contrasts with previous TMM research in stable teams, where interventions typically produce measurable convergence in mental models [22], [24]. The absence of significant effects aligns with theoretical propositions that shared cognition development requires sufficient time to process and integrate new information [27]. In temporary teams, this process may be particularly challenging due to limited shared history and the cognitive demands of simultaneously learning new information while applying it to complex tasks [14], [17].

Paired-samples t-tests showed no reliable pre-post change: Question 1a $t(15) = 0.212,\ p = .835$ and Question 4a $t(16) = 0.944,\ p = .358$ (Table 4). The corresponding independent-samples comparisons remained non-significant (all $|t| < 1.1,\ p > .30$) and the Time × Team interaction for the composite similarity score was likewise null, $F(1, 15) = 0.388,\ p = .543$ (Table 6). The Wilcoxon tests applied to within-group cohesion changes and the Mann-Whitney comparison of change scores between groups also remained non-significant. Taken together, these converging null findings indicate that the short briefing did not measurably align how team members conceptualized the incident, suggesting that the 15-minute threshold may be insufficient for meaningful TMM

convergence in *ad hoc* cybersecurity IR teams.

These results differ from findings in more stable team contexts, where even brief interventions have shown some effect on mental model similarity [26], [35]. For instance, Marks *et al.* [35] found that cross-training interventions produced significant improvements in team mental model similarity, but their interventions were longer and conducted with teams that had more opportunity for interaction. Similarly, Mathieu *et al.* [26] observed that mental model similarity developed over time through team interaction, suggesting that our 15-minute intervention may have been too brief to overcome the barriers to shared cognition in newly formed teams. The non-significant SBERT cohesion results also align with challenges noted by Mohammed *et al.* [22] regarding the measurement sensitivity required to detect subtle shifts in mental model convergence, particularly in early stages of team formation.

Future research should explore whether extending the intervention duration—perhaps to 30 or 45 minutes—might cross a critical threshold for TMM similarity development in *ad hoc* teams. Additionally, investigating staged interventions (e.g., brief initial training followed by structured reflection points during task execution) might reveal more effective approaches for accelerating mental model convergence in time-constrained settings. Methodologically, incorporating network analysis of team communication patterns alongside semantic similarity measures could provide more sensitive detection of emerging shared understanding before it manifests in explicit mental model measures [30].

6.3 Effects on Team Mental Model accuracy (SRQ2)

The theoretical foundation of Team Mental Model accuracy posits that teams with more accurate mental models—those that align with expert or "gold standard" conceptualizations—should perform more effectively in complex tasks [21], [24]. As Wulff *et al.* [95] emphasize, for scientific discoveries to be valid, a phenomenon must be accurately described, and its interpretation must withstand scrutiny by addressing appropriate counterfactuals and systematically eliminating competing explanations. In our study, it is attempted to enhance TMM accuracy through targeted cross-training, providing team members with insight into others' roles and responsibilities—a theoretically sound approach supported by previous research [35].

The checklist showed a descriptive gain of one point for the experimental team and a loss of half a point for the control team; the resulting Difference-in-Differences estimate was $+1.5$ points (15%), but no follow-up test reached significance (see Table 11). This pattern aligns with theoretical expectations that accuracy may be more responsive to

brief interventions than similarity [33], [34], as accuracy can improve through individual learning even before team-level convergence occurs. The descriptive improvement, while not statistically significant, suggests that the intervention may have initiated a positive trajectory in mental model accuracy that might become more pronounced with additional training or experience.

The differential response between the experimental and control teams warrants closer examination. The experimental team's modest gain in accuracy points to potential benefits of the TMM intervention for enhancing individual team members' understanding of correct response procedures. Conversely, the control team's slight decline in accuracy raises questions about potential interference effects when teams receive general security information rather than targeted TMM training. This pattern, though not statistically significant, is theoretically consistent with research suggesting that non-targeted information can sometimes create cognitive interference in complex task environments [27].

As a result, the data do not provide significant evidence that fifteen minutes of TMM training can materially sharpen a team's collective grasp of incident facts or required actions. However, the descriptive pattern suggests that accuracy improvements may precede similarity developments in the TMM formation process, consistent with theoretical models of team cognition development [22], [26]. While our results did not show statistically significant improvements in TMM accuracy following the intervention, this finding itself contributes to theoretical understanding by delineating boundary conditions for TMM development. The theoretical assumption that shared understanding can be rapidly developed through brief interventions may require qualification: the 15-minute timeframe tested in our study appears insufficient to meaningfully alter the accuracy of team members' mental models in the complex socio-technical domain of cybersecurity incident response.

These findings partially align with previous research on TMM accuracy in other domains. Edwards *et al.* [33] found that brief training interventions could improve the accuracy of team mental models in military teams, though their interventions were longer in duration. Similarly, Lim and Klein [34] observed that accuracy was sometimes more responsive to training than similarity, particularly in the early stages of team development. The modest descriptive improvement observed in our study, while not reaching statistical significance, suggests similar mechanisms may be at work in cybersecurity IR teams, albeit requiring more intensive intervention to reach significant levels of improvement.

An important methodological limitation emerged in our measurement strategy. Although team performance appeared strong based on observer notes, the written reports—used as our primary assessment tool—were consistently of lower quality. As Wulff *et al.*

observe, “measurement error in model predictors is one of the sources of endogeneity” that can undermine the validity of study outcomes [95]. This disparity between observed performance and written documentation raises concerns that our accuracy metric may have reflected participants’ writing proficiency rather than the actual quality of their team mental models, potentially masking the effects of the intervention. To address this issue, future studies should incorporate complementary measurement strategies, such as structured real-time observation protocols, to more reliably assess TMM accuracy.

Future research should investigate whether the descriptive accuracy improvements observed here might reach statistical significance with larger sample sizes or slightly longer interventions. Additionally, exploring differential effects of various intervention components (e.g., isolating the effects of cross-training versus strategy briefing) could reveal which elements most effectively enhance mental model accuracy in *ad hoc* teams. Incorporating process measures during task execution would also help clarify whether accuracy improvements manifest in actual decision-making behaviors before becoming detectable in post-task assessments.

6.4 Self-Reflection

Implementation constraints All exercise artifacts were presented in English, and the exercise itself was conducted in English. Although English is the official language of the company, a language barrier was still present, as only one of the eighteen participants was a native speaker. This linguistic heterogeneity introduces theoretical considerations regarding the interaction between language processing and cognitive load in TMM development [27]. When team members must simultaneously translate technical concepts and integrate new information into their mental models, the cognitive resources available for model development may be significantly reduced, potentially explaining some of the limited intervention effects observed.

Previous TMM research has rarely addressed multilingual contexts explicitly, representing a gap in the literature that our study helps illuminate. However, research on team cognition in multinational teams [80] suggests that linguistic diversity can create additional barriers to shared understanding. Our observations of language-related challenges align with these findings and extend them to the cybersecurity IR domain, where technical terminology adds another layer of complexity.

A concise form and a short pre-exercise tutorial could reduce cognitive load and missing data. Because outcome measurement relied on self-completed reports, behaviors that observers noted in real time were not always captured, underscoring the need for future

studies to integrate trained raters armed with behaviorally anchored check-lists. Future research should also consider developing and validating multilingual TMM assessment instruments or incorporating translation support to reduce language-related cognitive load. Additionally, studies might explicitly compare TMM development in linguistically homogeneous versus heterogeneous teams to isolate the effects of language diversity on intervention efficacy.

Upon closer examination of our intervention design, the language-related limitations underscore the significance of accounting for linguistic diversity when developing TMM training protocols. The effectiveness of the intervention may have been enhanced through the inclusion of visual elements, clearer and more accessible language, or multilingual support materials—each aimed at reducing the cognitive demands placed on non-native English speakers. This highlights a key consideration for both the methodological design of future studies and the practical implementation of TMM-based interventions within multinational organizational settings.

Scenario imbalance Scenario A offered a live log view that scaffolded timeline reconstruction, whereas Scenario B omitted this aid and included fewer hints, imposing greater cognitive demand. Such asymmetry risks ceiling effects in one condition and floor effects in the other, thereby obscuring treatment differences. From a theoretical perspective, this scenario imbalance interacts with the core TMM development processes we aimed to study. When cognitive resources are disproportionately allocated to basic information gathering (as in Scenario B), fewer resources remain available for the higher-order integration processes necessary for mental model development and alignment [27], [39]. This imbalance may have particularly affected the experimental team, who faced the more challenging scenario after receiving the TMM intervention, potentially masking intervention effects.

Previous experimental research on TMMs has emphasized the importance of equivalent task difficulty across conditions [26], [35]. Our experience aligns with methodological observations by Mohammed *et al.* [22], who noted that scenario characteristics can significantly influence the measurement of TMM constructs. The challenge of creating balanced scenarios while maintaining ecological validity represents an ongoing tension in TMM research, particularly in complex domains like cybersecurity where standardizing task difficulty is inherently challenging.

Independent pilot testing and iterative adjustment of scenario difficulty would help equalize cognitive load across conditions and yield cleaner contrasts. Future studies should consider employing expert panels to rate scenario difficulty across multiple dimensions (information

availability, technical complexity, time pressure) to ensure better calibration. Alternatively, within-subjects designs where teams encounter multiple scenarios of varying difficulty might help control for these effects, though such designs introduce their own challenges regarding practice effects and experimental duration.

A critical examination of our experimental design reveals that the scenario imbalance underscores a broader tension between ecological validity and experimental control in TMM research. Although real-world cybersecurity incidents inherently differ in complexity and information availability, experimental settings demand more precisely calibrated scenarios to reliably attribute effects to specific interventions. Future studies may benefit from hybrid designs that preserve ecological realism while introducing greater control—such as modular scenarios that can be dynamically adjusted in response to team performance to ensure a consistent level of challenge.

Team Composition and Leadership Effects Random assignment did not eliminate expertise asymmetry: the control-team leader had prior penetration-testing and incident-response experience, while the experimental-team leader had little relevant background and, according to facilitator notes, did not apply the newly taught techniques. Specifically, the control-team leader possessed years of experience in cybersecurity, including three years specializing in penetration testing, and had previously participated in four major incident response exercises.

In contrast, the experimental-team leader had a primarily management background with only basic security awareness training and no formal certifications in cybersecurity. More crucially, despite receiving training in specific methods such as strategic briefing techniques defined in our literature review, the experimental team leader failed to implement any of these approaches during the exercise. This implementation failure represents a fundamental disconnect between training and behavioral change, essentially nullifying a key aspect of our intervention. By definition, training aims to produce behavioral change; in this respect, our intervention failed at the leadership level despite apparent engagement during the training session itself. This leadership implementation gap likely contributed significantly to the lack of measurable TMM improvements in the experimental team.

This expertise asymmetry introduces important theoretical considerations regarding the role of leadership in TMM development. Research on team leadership suggests that leaders serve as cognitive anchors for their teams, particularly in novel or ambiguous situations [96]. When leaders possess domain expertise, they can facilitate more accurate mental model development among team members through guided sensemaking and expert direction.

Conversely, when leaders lack domain expertise or fail to model trained behaviors, team members may default to existing mental models rather than incorporating new frameworks [24]. The control team leader's expertise may have partially compensated for the lack of TMM intervention, while the experimental team leader's limited engagement with the training may have attenuated potential intervention benefits.

In addition, this leadership arrangement represents a potential outlier within our study design. André emphasizes that outliers must be treated with caution in experimental research, warning that "any outlier exclusion procedure that is not blind to the hypothesis that researchers want to test may result in inflated Type I errors." [97] In line with this guidance, we chose not to exclude the data related to the team leader but to explicitly acknowledge its potential influence as a contextual factor when interpreting our findings. Moreover, excluding the team leader from similarity calculations while still incorporating their input as the accuracy reference would compromise the internal consistency and validity of the measurement framework.

Previous research has demonstrated that leader characteristics significantly influence team cognition development. Zaccaro *et al.* [96] found that leader expertise can accelerate team mental model formation, while Burke *et al.* [42] observed that leaders who model desired behaviors facilitate faster adoption of shared frameworks. Our observations align with these findings, suggesting that leader expertise and engagement may moderate the effects of formal TMM interventions. This interaction between leadership and intervention efficacy represents an important consideration for TMM research that has received limited attention in the literature.

Reassessing our intervention design, the disparity in leader expertise underscores the importance of accounting for leadership dynamics in the development of team mental models. The effectiveness of the intervention may have been enhanced by incorporating leader-specific components, such as guidance on modeling and reinforcing the targeted behaviors. Moreover, ensuring that team leaders are actively engaged and visibly committed to the intervention appears to be a critical factor in shaping its overall impact. Future efforts should consider integrating tailored elements for leaders, along with implementation fidelity checks, to support more consistent and effective delivery across teams.

Session timing In addition to other influencing factors, the experimental group completed its scenario late in the afternoon, at which point clear signs of fatigue were observed. Although sessions were scheduled during regular business hours, teams were required to convene off-site, and this change in setting did not reduce participants' workload earlier in

the day. As a result, the emergence of fatigue in the experimental group was understandable. These circumstances likely compromised internal validity and reduced the effectiveness of the experimental manipulation. To address such issues, future studies should consider controlling for differences in expertise through stratified randomization or matched-pair designs to ensure comparable leadership capability across groups. Furthermore, counterbalancing the timing of sessions would help control for fatigue-related effects. Researchers may also explore the interaction between leader expertise and TMM-focused interventions by manipulating both variables within a factorial design framework.

Intervention design for behavioral change A significant methodological challenge emerged in our study regarding intervention fidelity. Despite receiving specific training in strategic briefing techniques and other TMM-enhancing methods, the experimental team leader did not implement these approaches during the post-intervention exercise. This implementation failure highlights a fundamental challenge in training-based interventions: as McCambridge *et al.* (2014) [98] note, behavioral interventions aim to produce change, but this change is contingent on participants' willingness and ability to apply the training.

Despite the experimental group's exposure to a 15-minute briefing on TMM strategies during the intervention phase, the incident commander's delegation style, utilization of communication aids and overall coordination behavior in Scenario B remained practically indistinguishable from baseline performance, a result that the Knowledge–Attitude–Behavior (KAB) model helps to illuminate. According to KAB theory, the acquisition of declarative or procedural knowledge constitutes only an initial cognitive shift whose practical value is realized only when it instigates concomitant changes in attitudes—such as the leader's belief in the efficacy of shared cognition—and in turn fortifies the self-efficacy and motivational states that support behavioral enactment [99], [100]. Meta-analytic and experimental evidence drawn from health education, organizational safety and cybersecurity contexts consistently shows that interventions limited to didactic information transfer rarely produce measurable behavioral change unless they are complemented by affective persuasion, repeated behavioral rehearsal, real-time feedback and environmental reinforcement [101], [102]. In the present study, the intervention delivered only the knowledge component: the leader received a lecture-style download with no opportunities to practice the recommended communication scripts, to receive formative coaching or to observe model performances that could recalibrate his attitudes toward collaborative command; moreover, no post-exercise fidelity checks or performance cues were embedded to support ongoing adoption. Implementation-science literature emphasizes that such fidelity drivers—continuous coaching, performance metrics and context-embedded prompts—are indispensable when behavioral change hinges on a single high-leverage actor, particularly

in time-pressured incident response settings where entrenched command habits prevail [103]. Absent these KAB-aligned supports, the leader's pre-existing attitudes and habitual behaviors likely dominated, leaving the cognitive gains inert and explaining the observed null effect on TMM similarity and incident-response performance.

Future research must include the TMM strategies as ground rules or procedures in the exercise, rather than a helper body of knowledge in the intervention. When there is no behavioral change on the experiment group after intervention, it was not reasonable to have a measurement result that does not reflect the effectiveness of the intervention. The TMM strategies must be listed as implementable instructions to be integrated, either as team rituals or structured communication methods.

Situational awareness lens on the pattern Endsley's three-level SA framework [39] clarifies the asymmetric outcome. Questions 1 and 4 target Level 1 perception, Questions 2 and 6 tap Level 2 comprehension and rudimentary projection, while the free-text question spans all SA levels. The only consistent descriptive gain surfaced at the factual-perception layer: participants in the experimental group recalled isolated incident indicators slightly better, but that gain did not propagate upward to shared comprehension or coordinated projection. This pattern aligns with theoretical models of situation awareness development, which suggest that perception (Level 1) typically develops before comprehension (Level 2) and projection (Level 3) [39]. It also corresponds to theoretical propositions about the sequential nature of TMM development, where factual alignment often precedes deeper conceptual integration [22], [26].

This finding parallels research by Cooke *et al.* [30], who observed that team training interventions often impact lower-level cognitive processes before affecting higher-order team cognition. Similarly, Endsley's work with aviation teams [39] demonstrated that situation awareness develops sequentially, with perception improvements preceding comprehension and projection gains. Our results extend these observations to the cybersecurity domain, suggesting similar cognitive development patterns despite the unique characteristics of IR contexts.

In other words, the intervention bolstered what teams noticed yet failed to transform that noticing into a convergent understanding of what the cues meant or what would happen next. Future research should investigate whether extended or repeated interventions might facilitate the progression from improved perception to enhanced comprehension and projection. Studies might also explore targeted interventions specifically designed to accelerate the development of Level 2 and Level 3 situation awareness in *ad hoc* teams,

potentially through structured sense-making exercises or guided scenario walkthroughs.

Reevaluating our intervention design, the variation in effects across situation awareness (SA) levels suggests that the 15-minute format may have been more effective in supporting perceptual processes than in fostering higher-order cognitive functions such as comprehension and projection. Future interventions could be strengthened by explicitly targeting all SA levels, incorporating dedicated elements to develop deeper understanding and forward-looking reasoning. Furthermore, refining the measurement approach to more precisely capture changes at each SA level may help identify subtle improvements that may have gone undetected with the current instruments.

Methodological Limitations The sample of eighteen individuals and two team-level checklists yielded less than forty per cent power to detect moderate effects, raising the risk of Type II error. This statistical limitation intersects with theoretical considerations regarding the measurement of team cognition constructs. TMM theory suggests that mental models are multidimensional, dynamic constructs that may require multiple measurement approaches to fully capture [22], [27]. Our limited sample size and measurement approach may have been insufficient to detect subtle but theoretically meaningful changes in these complex constructs, particularly given the brief nature of the intervention and the challenging task environment.

Our measurement approach relied primarily on written incident reports to assess TMM accuracy and similarity. While this method offered practical advantages in our experimental setting, it introduced significant limitations. As observed during the exercise, teams demonstrated more sophisticated coordination and understanding than was reflected in their written reports. This created a noticeable disconnect between actual team performance and the primary measurement instrument. Wulff*et al.* [95] highlight that "knowledge of performance may also bias how observers rate the [leadership]," pointing to the complex relationship between actual performance and its documentation. In our case, the inverse may have occurred: teams exhibited effective real-time performance, yet failed to adequately capture this in written form, resulting in measurement error that may have obscured the effects of the intervention. Moreover, this discrepancy raises specific concerns about construct validity—whether our written-report instrument truly measured the cognitive constructs associated with TMMs, or instead reflected unrelated factors such as writing proficiency, documentation habits, or motivation to complete the report. This methodological issue underscores the importance of adopting multi-method assessment strategies in TMM research, particularly in high-tempo environments where written documentation may be deprioritized or inconsistently executed.

Mono-method bias persists because the Likert items were purpose-built and lack external validation, and language dependence further limits generalisability. These methodological challenges echo concerns raised by Mohammed *et al.* [22] regarding the measurement of TMM constructs. Their comprehensive review highlighted the importance of using multiple, validated measurement approaches to capture the multidimensional nature of team cognition. Similarly, Cooke *et al.* [27] emphasized the need for ecologically valid measurement approaches that capture both explicit and implicit aspects of team knowledge. Our study's limitations align with these methodological challenges, which remain persistent issues in TMM research.

These constraints counsel caution in extrapolating conclusions beyond the present setting. Future research should address these limitations through larger samples, validated measurement instruments, and mixed-method approaches that triangulate findings across multiple data sources. Developing and validating domain-specific TMM measurement instruments for cybersecurity IR would be particularly valuable for advancing research in this area. Additionally, incorporating physiological or behavioral measures alongside self-report instruments might provide more sensitive detection of TMM development, particularly in early stages.

Reevaluating our methodological approach, the identified limitations illustrate the inherent difficulty of investigating complex cognitive constructs within ecologically valid settings. While laboratory-based studies may provide greater statistical power through larger sample sizes, they often do so at the expense of the contextual nuance that defines real-world incident response environments. Future research could benefit from innovative study designs that seek to balance these trade-offs—such as employing multiple case studies with standardized measurement frameworks or utilizing simulation-based methodologies that support larger participant groups while preserving situational authenticity.

Practical implications Even a micro-intervention can nudge factual recall upward, but achieving genuine mental model convergence appears to require more extensive practice, structured debriefing, or guided role rotation. This observation aligns with theoretical models of expertise development and skill acquisition [104], which suggest that complex cognitive skills typically require deliberate practice and feedback to develop fully. In the context of TMM theory, our findings suggest that different aspects of team cognition may develop at different rates and through different mechanisms [22], [26], with factual knowledge responding more quickly to brief interventions than deeper conceptual alignment.

Organizations that routinely assemble temporary incident-response cells may benefit from pairing short briefings with observer-based feedback loops if the goal is to forge a common cognitive picture rather than merely improve individual fact retention. This recommendation aligns with research by Marks *et al.* [35], who found that team debriefs significantly enhanced the effectiveness of cross-training interventions. Similarly, Mathieu *et al.* [24] observed that feedback mechanisms accelerated the development of shared mental models in aviation teams. Our findings extend these observations to cybersecurity contexts, suggesting similar principles apply despite the unique characteristics of IR environments.

Future implementations could explore integrated approaches that combine brief initial training with structured reflection points during incident response. Organizations might also consider developing standardized TMM support tools that can be rapidly deployed during incidents, such as shared visualization platforms or structured communication protocols. Evaluating the effectiveness of such integrated approaches would provide valuable insights for both research and practice.

Reflecting on the practical implications of our findings, the modest impact of the brief intervention underscores the ongoing tension between operational constraints and the cognitive demands of effective team coordination in cybersecurity incident response. Although organizations typically face strict time limitations for team preparation during incidents, our results indicate that even modest increases in preparatory effort and sustained support may contribute meaningfully to team performance. Identifying the appropriate balance between time investment and cognitive gains remains a key consideration for both future research and practical implementation in organizational settings.

6.5 Directions for Further Inquiry

Future work is advised to employ larger samples, balanced scenarios, and multi-method assessments that fuse questionnaires with behavioral observation and communication analysis. From a theoretical perspective, such methodological refinements would help address fundamental questions about the development trajectory of TMMs in *ad hoc* teams [22], [27]. By capturing both explicit and implicit aspects of team cognition across multiple time points, future research could clarify the mechanisms through which shared understanding develops in temporary teams and identify critical thresholds for intervention efficacy. Experiments that vary the briefing dosage—from the present 15-minute primer to $\sim$ 30-minute or staged brief-plus-reflection formats—may help pinpoint the minimum viable intervention for TMM convergence.

Given the significant limitations we encountered with report-based measurement, future research should specifically incorporate trained observers into the experimental design. These observers should use standardized behavioral markers and interaction pattern coding to directly assess team cognition manifestations during task execution, rather than relying primarily on post-hoc self-reports. Observer-based measurement would be particularly valuable for capturing the quality of team coordination, information sharing patterns, and decision-making processes that may reflect shared mental models but fail to appear in written reports. Additionally, recording and analyzing team communications could provide more direct evidence of mental model convergence or divergence as it occurs in real-time, offering insights that our report-based approach could not capture. In parallel, researchers should test workflow-embedded scaffolds—e.g., automated role prompts or checklist pop-ups inside chat-ops and SOAR tools—to evaluate whether digital supports can substitute for longer classroom training. Moreover, manipulating linguistic load by comparing English-only materials with simplified or fully translated versions would clarify how language barriers and cognitive load jointly affect TMM formation.

Additionally, future studies should directly address the disconnect between observed team performance and documentation quality by implementing dual measurement approaches. This could include real-time performance metrics captured during the exercise alongside traditional report-based assessments, allowing researchers to quantify the gap between actual performance and its documentation. Such an approach would help clarify whether non-significant findings reflect true intervention limitations or measurement artifacts arising from documentation challenges in high-pressure environments.

Additionally, researchers are encouraged to report results both with and without outliers, as recommended by Wulff *et al.* [95], in order to provide transparency regarding the influence of exceptional cases—such as the control team leader in our study. In retrospect, our experimental design did not anticipate the potential impact of outliers and incorporated a strong dependence on team leaders for coordinating the exercise. Future studies should pre-register explicit outlier-handling rules and complement classical analyses with robust statistics (e.g., trimmed means or bootstrapped confidence intervals) to mitigate undue leverage. As a result, the presence of an outlier in one of these leadership roles had a disproportionate effect on both team dynamics and outcome measures. Design-wise, building redundancy into critical roles—such as appointing a deputy leader or rotating the incident-commander function—will help ensure that single exceptional performers neither mask nor exaggerate intervention effects.

Longitudinal follow-up would clarify whether brief gains in perception decay or consolidate, and rigorous treatment-fidelity checks are likely to be important when team leaders

are expected to model the trained techniques. Future designs should also target all three levels of situation awareness, assessing whether perception gains cascade into comprehension and projection under extended or repeated interventions. This recommendation builds on research by Mathieu *et al.* [26], who observed that mental models evolve over time through team interaction and experience. Similarly, Burke *et al.* [42] emphasized the importance of implementation fidelity in team training interventions, particularly regarding leadership behaviors. Our suggestions extend these insights to the specific challenges of cybersecurity IR contexts, where time constraints and high stakes create unique conditions for TMM development. To specifically address intervention implementation challenges, future research should incorporate structured implementation protocols with explicit behavioral checkpoints for team leaders. These might include pre-briefing commitment contracts, mid-exercise coaching interventions, or post-training behavioral rehearsal to increase the likelihood that trained techniques are actually implemented. This approach would help distinguish between intervention ineffectiveness (where properly implemented techniques fail to produce effects) and implementation failure (where techniques are never properly applied).

Beyond methodological refinements, future research should explore several theoretical questions raised by our findings. First, studies might investigate the differential development of various TMM content domains in cybersecurity contexts, examining whether task-related, team-related, or technology-related mental models respond differently to brief interventions. Second, research could explore the interaction between individual expertise and team-level cognition, clarifying how diverse knowledge backgrounds influence TMM development in *ad hoc* teams. Finally, studies might examine how different intervention components (e.g., cross-training, strategy briefing, role rotation) contribute to TMM development, potentially identifying more efficient approaches for accelerating shared understanding in time-constrained settings.

6.6 Summary

Across SRQ 1 and SRQ 2, the evidence indicates that the main research question was *not* supported: a 15-minute TMM briefing produced a modest rise in factual accuracy but no detectable gain in mental-model similarity. This result refines theory in three ways. First, it implies that 15 minutes may approach the lower boundary for effective TMM work in *ad hoc* cybersecurity teams. Second, it suggests that elements of team cognition develop at different speeds—surface-level facts may adjust quickly, whereas deeper conceptual alignment lags [22], [26]. Third, it reinforces calls to account for leadership, expertise distribution, and task complexity when translating TMM theory to new domains [27].

Several factors probably muted the intervention's impact: language burden, complex reporting requirements, scenario imbalance, leader-experience disparities, and end-of-day fatigue. These challenges echo those identified in other ecologically valid TMM studies [22], [96]. Compared with work on stable military or aviation teams [24], [35], our findings hint that temporary response cells may require either longer or differently structured interventions.

The study nonetheless contributes methodologically by combining semantic similarity analysis with traditional surveys, extending the TMM toolset for capturing multidimensional cognition.

Two limitations warrant emphasis. First, the experimental team leader did not apply the briefing strategies, exposing a fidelity gap. Future designs will likely need accountability checks to ensure leaders model the intervention. Second, written reports proved a blunt instrument for latent cognition; the disconnect between team behavior and report quality raises concerns about construct validity.

Addressing these issues—through streamlined instruments, balanced scenarios, observer ratings, stronger fidelity controls, and workflow-embedded cognitive scaffolds—could clarify the true value of TMM briefings. Embedding TMM principles directly into incident-management platforms may also help organizations that lack time for extensive pre-incident training. Progress will depend on transparent, pre-registered studies that blend theoretical rigor with operational viability.

In sum, although the briefing did not yield statistically significant gains, the study advances understanding of rapid TMM development in high-stakes, temporary teams and illustrates a mixed-method evaluation framework for future investigations.

7. Conclusion

This thesis investigated ways to strengthen the cognitive foundations of *ad hoc* cybersecurity incident-response teams. Drawing on research from multiple areas of research, it tested whether a brief Team Mental Model (TMM) training (team intervention) could improve shared understanding and coordination.

Eighteen volunteers were randomly assigned to two matched teams that tackled pre- and post-intervention tabletop scenarios. Reduced bias is ensured by the quasi-experiment: a control and an experimental group tested pre- and post-training scenarios. Control group took a pseudo-training while experimental group took the TMM-focused training material. TMM similarity and accuracy were measured with three tools: Likert items, SBERT-based semantic indices, and a domain-expert checklist. This mixed-method design offers a reusable template for future team-cognition studies.

The training, unfortunately, did not produce statistically significant gains in TMM similarity or accuracy, although the experimental group showed a small descriptive increase in factual recall. These results suggest that TMM-specific trainings may raise individual cue recognition but are unlikely, by themselves, to create the deeper shared models needed for rapid, coordinated action. Contextual factors such as language load, scenario imbalance, and uneven expertise, probably diluted any treatment effect.

Even with non-significant outcomes, the study contributes empirical evidence on the lower limits of TMM interventions for temporary teams. It also demonstrates a practical measurement bundle that combines narrative and survey data.

Two limitations deserve emphasis, out of many others mentioned in the Chapter 6 Discussion: First, the experimental group team leader failed to implement the training strategies, highlighting a gap between the design and implementation; future studies should include accountability checks. When experiment group does not implement the requested strategies, measurement of the outcome does not provide value. Second, written reports captured performance poorly, questioning construct validity; observer ratings or live-feed analytics may offer richer data.

While the theoretical background, experiment design and metrics were successful, the unforeseen implementation issues affected the outcomes. Therefore, the study is an

artifact of lessons learned in the applied research area. Researchers might now test longer or repeated trainings, add discourse-analytic pipelines, and run multi-site replications across languages and organizations. Embedding TMM scaffolds directly into incident-management platforms could help teams that lack time for extensive pre-incident training. By outlining both the promise and the limits of micro-duration TMM training, this work provides a springboard for improving the cognitive readiness of *ad hoc* IR teams.

References

[1] E. Network and I. S. Agency, *ENISA threat landscape 2024 : July 2023 to June 2024*, A. Malatras, M. Theocharidou, I. Lella, and E. Tsekmezoglou, Eds. European Network and Information Security Agency, Sep. 2024.

[2] F. IC3, "2023 internet crime report," FBI, Tech. Rep., 2024.

[3] Mandiant, "Mandiant m-trends 2025 report," Google, Tech. Rep., 2025.

[4] Cybersecurity and Infrastructure Security Agency (CISA), National Security Agency (NSA), Federal Bureau of Investigation (FBI), U.S. Department of Energy (DOE), Environmental Protection Agency (EPA), *et al.*, "Joint guidance: Identifying and mitigating living off the land techniques," Cybersecurity and Infrastructure Security Agency (CISA), Tech. Rep., Feb. 2024, TLP:CLEAR. [Online]. Available: https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-144a.

[5] W. contributors, *Software as a service — Wikipedia, the free encyclopedia*, [Online; accessed 11-May-2025], 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Software_as_a_service&oldid=1284894122.

[6] P. V. Falade, "Decoding the threat landscape: Chatgpt, fraudgpt, and wormgpt in social engineering attacks," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 185–198, Oct. 2023, ISSN: 2456-3307. DOI: 10.32628/cseit2390533.

[7] S. R. Department, *Estimated annual cost of cybercrime worldwide*, 2023.

[8] I. Security and P. Institute, "Cost of a data breach report 2024," IBM Security and Ponemon Institute, Tech. Rep., 2024.

[9] W. E. Forum, "Global cybersecurity outlook 2025," World Economic Forum, Tech. Rep., 2025.

[10] Axelos, *ITIL Foundation*. Norwich, England: Stationery Office Books, Feb. 2019.

[11] M. Nieles, K. Dempsey, and V. Y. Pillitteri, *An introduction to information security*. National Institute of Standards and Technology, Jun. 2017. DOI: 10.6028/nist.sp.800-12r1.

[12] N. I. o. S. Computer Security Division Information Technology Laboratory and Technology, *Minimum security requirements for federal information and information systems*. National Institute of Standards and Technology (U.S.), 2006. DOI: 10.6028/nist.fips.200.

[13] A. Nelson, *Incident Response Recommendations and Considerations for Cybersecurity Risk Management:: A CSF 2.0 Community Profile*. National Institute of Standards and Technology, 2025. DOI: 10.6028/nist.sp.800-61r3.

[14] R. Kleij, G. Kleinhuis, and H. Young, "Computer security incident response team effectiveness: A needs assessment," *Frontiers in Psychology*, vol. 8, no. DEC, 2017.

[15] E. A. Skryabina, N. Betts, G. Reedy, P. Riley, and R. Amlôt, "The role of emergency preparedness exercises in the response to a mass casualty terrorist incident: A mixed methods study," *International Journal of Disaster Risk Reduction*, vol. 46, p. 101 503, 2020, ISSN: 2212-4209. DOI: https://doi.org/10.1016/j.ijdrr.2020.101503.

[16] P. Shedden, A. Ahmad, and A. Ruighaver, "Organisational learning and incident response: Promoting effective learning through the incident response process," *Australian Information Security Management Conference*, Mar. 2012.

[17] R. Bakker, "Taking stock of temporary organizational forms: A systematic review and research agenda," *International Journal of Management Reviews*, vol. 12, pp. 466–486, 2010.

[18] M. Hällgren and E. Maaninen-Olsson, "Deviations and the breakdown of project management principles," *International Journal of Managing Projects in Business*, vol. 2, no. 1, pp. 53–69, Jan. 2009, ISSN: 1753-8378. DOI: 10.1108/17538370910930518.

[19] J. C. Shylanski and C. E. Wilke, "3.2.1 task teams – meeting the organizational challenges of systems engineering," *INCOSE International Symposium*, vol. 8, no. 1, pp. 236–243, Jul. 1998, ISSN: 2334-5837. DOI: 10.1002/j.2334-5837.1998.tb00034.x.

[20] R. Bhaskar, "A proposed integrated framework for coordinating computer security incident response team," *Journal of Information Privacy and Security*, vol. 1, no. 3, pp. 3–17, 2005.

[21] J. A. Cannon-Bowers, E. Salas, and S. Converse, "Shared mental models in expert team decision making.," *Lawrence Erlbaurm*, pp. 221–246, Jan. 1993.

[22] S. Mohammed, L. Ferzandi, and K. Hamilton, "Metaphor no more: A 15-year review of the team mental model construct," *Journal of Management*, vol. 36, no. 4, pp. 876–910, 2010. DOI: 10.1177/0149206309356804.

[23] S. Wise, C. Duffield, M. Fry, and M. Roche, "A team mental model approach to understanding team effectiveness in an emergency department: A qualitative study," *Journal of Health Services Research & Policy*, vol. 27, no. 1, pp. 14–21, Jul. 2021, ISSN: 1758-1060. DOI: 10.1177/13558196211031285.

[24] J. Mathieu, T. Heffner, G. Goodwin, E. Salas, and J. Cannon-Bowers, "The influence of shared mental models on team process and performance," *Journal of Applied Psychology*, vol. 85, pp. 273–283, Apr. 2000. DOI: 10.1037/0021-9010.85.2.273.

[25] J. Mathieu, M. T. Maynard, T. Rapp, and L. Gilson, "Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future," *Journal of management*, vol. 34, no. 3, pp. 410–476, 2008.

[26] J. E. Mathieu, T. S. Heffner, G. F. Goodwin, J. A. Cannon-Bowers, and E. Salas, "Scaling the quality of teammates' mental models: Equifinality and normative comparisons," *Journal of Organizational Behavior*, vol. 26, no. 1, pp. 37–56, 2005. DOI: 10.1002/job.296.

[27] N. J. Cooke, E. Salas, J. A. Cannon-Bowers, and R. J. Stout, "Measuring team knowledge," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 42, no. 1, pp. 151–173, Mar. 2000, ISSN: 1547-8181. DOI: 10.1518/001872000779656561.

[28] S. Mohammed and B. C. Dumville, "Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries," *Journal of Organizational Behavior*, vol. 22, no. 2, pp. 89–106, Mar. 2001, ISSN: 1099-1379. DOI: 10.1002/job.86.

[29] I. L. Janis, *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.* Houghton Mifflin, 1972.

[30] N. J. Cooke, P. A. Kiekel, E. Salas, R. Stout, C. Bowers, and J. Cannon-Bowers, "Measuring team knowledge: A window to the cognitive underpinnings of team performance.," *Group Dynamics: Theory, Research, and Practice*, vol. 7, no. 3, pp. 179–199, Sep. 2003, ISSN: 1089-2699. DOI: 10.1037/1089-2699.7.3.179.

[31] M. Hällgren, "Groupthink in temporary organizations," *International Journal of Managing Projects in Business*, vol. 3, no. 1, pp. 94–110, Jan. 2010, ISSN: 1753-8378. DOI: 10.1108/17538371011014044.

[32] S. van der Haar, "Getting on the same page: Team learning and team cognition in emergency management command-and-control teams," Ph.D. dissertation, Education, Child Studies, Faculty of Social, and Behavioural Sciences, Leiden University, 2014.

[33] B. D. Edwards, E. A. Day, W. Arthur Jr, and S. T. Bell, "Relationships among team ability composition, team mental models, and team performance.," *Journal of applied psychology*, vol. 91, no. 3, p. 727, 2006.

[34] B.-C. Lim and K. J. Klein, "Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy," *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, vol. 27, no. 4, pp. 403–418, 2006.

[35] M. A. Marks, J. E. Mathieu, and S. J. Zaccaro, "A temporally based framework and taxonomy of team processes," *The Academy of Management Review*, vol. 26, no. 3, p. 356, Jul. 2001, ISSN: 0363-7425. DOI: 10.2307/259182.

[36] C. C. Bowers Clint A.;Braun, "Team workload: Its meaning and measurement," in *Team performance assessment and measurement: Theory, methods, and applications*, B. B. Morgan Jr., Ed. Lawrence Erlbaum Associates Publishers, 1997, pp. 85–108.

[37] R. Rico, M. Sánchez-Manzanares, F. Gil, and C. Gibson, "Team implicit coordination processes: A team knowledge–based approach," *Academy of Management Review*, vol. 33, no. 1, pp. 163–184, 2008. DOI: 10.5465/amr.2008.27751276. eprint: https://doi.org/10.5465/amr.2008.27751276.

[38] A. A. Nofi, "Defining and measuring shared situational awareness," *Center for Naval Analyses*, p. 74, Nov. 2000.

[39] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems.," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[40] M. T. Maynard, D. M. Kennedy, and S. A. Sommer, "Team adaptation: A fifteen-year synthesis (1998–2013) and framework for how this literature needs to "adapt" going forward," *European Journal of Work and Organizational Psychology*, vol. 24, no. 5, pp. 652–677, Jan. 2015, ISSN: 1464-0643. DOI: 10.1080/1359432x.2014.1001376.

[41] C. J. Resick, M. W. Dickson, J. K. Mitchelson, L. K. Allison, and M. A. Clark, "Team composition, cognition, and effectiveness: Examining mental model similarity and accuracy.," *Group Dynamics: Theory, Research, and Practice*, vol. 14, no. 2, p. 174, 2010.

[42] C. S. Burke, K. C. Stagl, E. Salas, L. Pierce, and D. Kendall, "Understanding team adaptation: A conceptual analysis and model.," *Journal of Applied Psychology*, vol. 91, no. 6, pp. 1189–1207, Nov. 2006, ISSN: 0021-9010. DOI: 10.1037/0021-9010.91.6.1189.

[43] M. A. M. John E. Mathieu and S. J. Zaccaro, "Multiteam systems," in *Handbook of Industrial, Work and Organizational Psychology, Volume 2 Organziational Psychology*, H. K. S. Neil Anderson Deniz S. Ones and C. Viswesvaran, Eds. SAGE Publications, 2001, pp. 289–131.

[44] E. Salas, D. E. Sims, and C. S. Burke, "Is there a "big five" in teamwork?" *Small Group Research*, vol. 36, no. 5, pp. 555–599, Oct. 2005, ISSN: 1552-8278. DOI: 10.1177/1046496405277134.

[45] B. H. Johnsen, R. Espevik, J. Eid, Ø. Østerås, J. K. Jacobsen, and G. Brattebø, "Coordinating mechanisms are more important than team processes for geographically dispersed emergency dispatch and paramedic teams," *Frontiers in Psychology*, vol. 13, Mar. 2022, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2022.754855.

[46] L. A. DeChurch and J. R. Mesmer-Magnus, "The cognitive underpinnings of effective teamwork: A meta-analysis.," *Journal of Applied Psychology*, vol. 95, no. 1, pp. 32–53, Jan. 2010, ISSN: 0021-9010. DOI: 10.1037/a0017328.

[47] K. A. Smith-Jentsch, J. E. Mathieu, and K. Kraiger, "Investigating linear and interactive effects of shared mental models on safety and efficiency in a field setting.," *Journal of applied psychology*, vol. 90, no. 3, p. 523, 2005.

[48] F. R. H. Zijlstra, M. J. Waller, and S. I. Phillips, "Setting the tone: Early interaction patterns in swift-starting teams as a predictor of effectiveness," *European Journal of Work and Organizational Psychology*, vol. 21, no. 5, pp. 749–777, Oct. 2012, ISSN: 1464-0643. DOI: 10.1080/1359432x.2012.690399.

[49] G. Grote, "Leading high-risk teams in aviation," in *Leadership Lessons from Compelling Contexts*. Emerald Group Publishing Limited, Mar. 2016, pp. 189–208. DOI: 10.1108/s1479-357120160000008006.

[50] T. Bijlsma and C. Broers, "Crew resource management as shield and spear for safety and security," in T.M.C. Asser Press, The Hague, Jul. 2016, pp. 275–290, ISBN: 978-94-6265-134-0. DOI: 10.1007/978-94-6265-135-7_14.

[51] R. L. Helmreich, A. C. Merritt, and J. A. Wilhelm, "The evolution of crew resource management training in commercial aviation," *The International Journal of Aviation Psychology*, vol. 9, no. 1, pp. 19–32, Jan. 1999, ISSN: 1532-7108. DOI: 10.1207/s15327108ijap0901_2.

[52] T. Driskell, G. Funke, M. T. Tolston, A. Capiola, and J. Driskell, "Supporting fluid teams: A research agenda," *Frontiers in Psychology*, vol. Volume 15 - 2024, 2024, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2024.1327885.

[53] T. Driskell, G. Funke, M. Tolston, A. Capiola, and J. Driskell, "Composition considerations for fluid teams: A review," *Frontiers in Psychology*, vol. Volume 15 - 2024, 2024, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2024.1302022.

[54] B. A. A. White, A. Eklund, T. McNeal, A. Hochhalter, and A. C. Arroliga, "Facilitators and barriers to ad hoc team performance," *Baylor University Medical Center Proceedings*, vol. 31, no. 3, pp. 380–384, May 2018, ISSN: 1525-3252. DOI: 10.1080/08998280.2018.1457879.

[55] N. K. Roberts, R. G. Williams, C. J. Schwind, J. A. Sutyak, C. McDowell, D. Griffen, *et al.*, "The impact of brief team communication, leadership and team behavior training on ad hoc team performance in trauma care settings," *The American Journal of Surgery*, vol. 207, no. 2, pp. 170–178, Feb. 2014, ISSN: 0002-9610. DOI: 10.1016/j.amjsurg.2013.06.016.

[56] Institute of Medicine and Committee on Quality of Health Care in America, *To Err Is Human: Building a safer health system*, en. Washington, D.C., DC: National Academies Press, Mar. 2000.

[57] M. J. Burtscher and T. Manser, "Team mental models and their potential to improve teamwork and safety: A review and implications for future research in healthcare," *Safety science*, vol. 50, no. 5, pp. 1344–1354, 2012.

[58] E. Salas, D. R. Nichols, and J. E. Driskell, "Testing three team training strategies in intact teams: A meta-analysis," *Small Group Research*, vol. 38, no. 4, pp. 471–488, Aug. 2007, ISSN: 1552-8278. DOI: 10.1177/1046496407304332.

[59] M. A. Marks, S. J. Zaccaro, and J. E. Mathieu, "Performance implications of leader briefings and team-interaction training for team adaptation to novel environments.," *Journal of Applied Psychology*, vol. 85, no. 6, pp. 971–986, 2000, ISSN: 0021-9010. DOI: 10.1037/0021-9010.85.6.971.

[60] C. N. Lacerenza, S. L. Marlow, S. I. Tannenbaum, and E. Salas, "Team development interventions: Evidence-based approaches for improving teamwork.," *American Psychologist*, vol. 73, no. 4, pp. 517–531, May 2018, ISSN: 0003-066X. DOI: 10.1037/amp0000295.

[61] K. A. Smith-Jentsch, J. A. Cannon-Bowers, S. I. Tannenbaum, and E. Salas, "Guided team self-correction: Impacts on team mental models, processes, and effectiveness.," *Small Group Research*, vol. 39, no. 3, pp. 303–327, 2008.

[62] J. A. Cannon-Bowers and E. Salas, "Making decisions under stress: Implications for individual and team training.," in *Making decisions under stress: Implications for individual and team training.* American Psychological Association, 1998, pp. 17–38.

[63] C. Volpe, J. A. Cannon-Bowers, and E. Salas, "The impact of cross-training on team functioning: An empirical investigation.," *Human Factors*, vol. 38, no. 1, pp. 87–100, 1996. DOI: 10.1518/001872096778940741.

[64] E. Blickensderfer, J. A. Cannon-Bowers, and E. Salas, "Cross-training and team performance.," in *Making decisions under stress: Implications for individual and team training.* American Psychological Association, 1998, pp. 299–311, ISBN: 1557985251. DOI: 10.1037/10278-011.

[65] R. Espevik, B. H. Johnsen, and J. Eid, "Outcomes of shared mental models of team members in cross training and high-intensity simulations," *Journal of Cognitive Engineering and Decision Making*, vol. 5, no. 4, pp. 352–377, Nov. 2011, ISSN: 1555-3434. DOI: 10.1177/1555343411424695.

[66] M. A. Marks, M. J. Sabella, C. S. Burke, and S. J. Zaccaro, "The impact of cross-training on team effectiveness.," *Journal of Applied Psychology*, vol. 87, no. 1, pp. 3–13, 2002, ISSN: 0021-9010. DOI: 10.1037/0021-9010.87.1.3.

[67] R. Fernandez, S. Shah, E. D. Rosenman, S. W. J. Kozlowski, S. H. Parker, and J. A. Grand, "Developing team cognition: A role for simulation," *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, vol. 12, no. 2, pp. 96–103, Apr. 2017, ISSN: 1559-2332. DOI: 10.1097/sih.0000000000000200.

[68] E. Salas, J. L. Wildman, and R. F. Piccolo, "Using simulation-based training to enhance management education," *Academy of Management Learning & Education*, vol. 8, no. 4, pp. 559–573, Dec. 2009, ISSN: 1944-9585. DOI: 10.5465/amle.8.4.zqr559.

[69] R. J. Stout, J. A. Cannon-Bowers, and E. Salas, "The role of shared mental models in developing team situational awareness: Implications for training," in *Situational awareness*, Routledge, 2017, pp. 287–318.

[70] R. Tesler, S. Mohammed, K. Hamilton, V. Mancuso, and M. McNeese, "Mirror, mirror: Guided storytelling and team reflexivity's influence on team mental models," *Small Group Research*, vol. 49, no. 3, pp. 267–305, Aug. 2017, ISSN: 1552-8278. DOI: 10.1177/1046496417722025.

[71] J. J. van Rensburg, C. M. Santos, S. B. de Jong, and S. Uitdewilligen, "The five-factor perceived shared mental model scale: A consolidation of items across the contemporary literature," *Frontiers in Psychology*, vol. 12, Jan. 2022, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.784200.

[72] M. J. Waller, N. Gupta, and R. C. Giambatista, "Effects of adaptive behaviors and shared mental models on control crew performance," *Management Science*, vol. 50, no. 11, pp. 1534–1544, Nov. 2004, ISSN: 1526-5501. DOI: 10.1287/mnsc.1040.0210.

[73] S. Uitdewilligen, M. J. Waller, R. A. Roe, and P. Bollen, "The effects of team mental model complexity on team information search and performance trajectories," *Group & Organization Management*, vol. 48, no. 3, pp. 755–789, 2023. DOI: 10.1177/10596011211023219. eprint: https://doi.org/10.1177/10596011211023219.

[74] J. Van den Berg, "A basic set of mental models for understanding and dealing with the cyber-security challenges of today," *Journal of Information Warfare*, vol. 19, no. 1, pp. 26–47, 2020.

[75] R. Murimi, S. Blanke, and R. Murimi, "A decade of development of mental models in cybersecurity and lessons for the future," in *Proceedings of the International Conference on Cybersecurity, Situational Awareness and Social Media*. Springer Nature Singapore, 2023, pp. 105–132, ISBN: 9789811964145. DOI: 10.1007/978-981-19-6414-5_7.

[76] R. Floodeen, J. Haller, and B. Tjaden, "Identifying a shared mental model among incident responders," in *2013 Seventh International Conference on IT Security Incident Management and IT Forensics*, IEEE, 2013, pp. 15–25.

[77] J. Steinke, B. Alaybek, L. Fletcher, V. Wang, A. Tomassetti, K. Repchick, *et al.*, "Improving cybersecurity incident response team effectiveness using teams-based research," *IEEE Security & Privacy*, vol. 13, pp. 20–29, Jul. 2015. DOI: 10.1109/MSP.2015.71.

[78] J. Maier, "Mental models of cyber security attacks and their influence on the design of cyber security dashboards," Master's thesis, Technical University of Munich, Munich, May 2016.

[79] K. Kullman, L. Buchanan, A. Komlodi, and D. Engel, "Mental model mapping method for cybersecurity," in *HCI for Cybersecurity, Privacy and Trust*. Springer International Publishing, 2020, pp. 458–470, ISBN: 9783030503093. DOI: 10.1007/978-3-030-50309-3_30.

[80] J. E. Driskell, E. Salas, and T. Driskell, "Foundations of teamwork and collaboration.," *American Psychologist*, vol. 73, no. 4, pp. 334–348, May 2018, ISSN: 0003-066X. DOI: 10.1037/amp0000241.

[81] S. Uitdewilligen and M. J. Waller, "Information sharing and decision-making in multidisciplinary crisis management teams," *Journal of Organizational Behavior*, vol. 39, no. 6, pp. 731–748, Jun. 2018, ISSN: 1099-1379. DOI: 10.1002/job.2301.

[82] T. Grance, T. Nolan, K. Burke, R. Dudley, G. White, and T. Good, *Guide to test, training, and exercise programs for IT plans and capabilities*. National Institute of Standards and Technology, 2006. DOI: 10.6028/nist.sp.800-84.

[83] A. R. Martinez, "The role of shared mental models in team coordination crew resource management skills of mutual performance monitoring and backup behaviors," Doctoral Dissertation, The University of Southern Mississippi, 2015.

[84] E. Salas, K. C. Stagl, C. S. Burke, and G. F. Goodwin, "Fostering team effectiveness in organizations: Toward an integrative theoretical framework," en, *Nebr Symp Motiv*, vol. 52, pp. 185–243, 2007.

[85] J. C. Gorman, N. J. Cooke, and P. G. Amazeen, "Training adaptive teams," en, *Hum Factors*, vol. 52, no. 2, pp. 295–307, Apr. 2010. DOI: 10.1177/0018720810371689.

[86] P. C. I. (S. S. Council, *Glossary — pcisecuritystandards.org*, https://www.pcisecuritystandards.org/glossary/, [Accessed 11-05-2025], 2025.

[87] Sygnia, *Breaking the Virtual Barrier: From Web-Shell to Ransomware — sygnia.co*, https://www.sygnia.co/threat-reports-and-advisories/breaking-the-virtual-barrier-web-shell-to-ransomware/, [Accessed 11-05-2025], 2025.

[88] J. L. Wildman, E. Salas, and C. P. R. Scott, "Measuring cognition in teams: A cross-domain review," *Human Factors*, vol. 56, no. 5, pp. 911–941, 2014, PMID: 25141596. DOI: 10.1177/0018720813515907. eprint: https://doi.org/10.1177/0018720813515907.

[89] J. Team, *Jasp (version 0.19.3)[computer software]*, 2025. [Online]. Available: https://jasp-stats.org/.

[90] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge, May 2013, ISBN: 9781134742707. DOI: 10.4324/9780203771587.

[91] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas," *Frontiers in Psychology*, vol. 4, 2013, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00863.

[92] L. Wilkinson, "Statistical methods in psychology journals: Guidelines and explanations.," *American Psychologist*, vol. 54, no. 8, pp. 594–604, Aug. 1999, ISSN: 0003-066X. DOI: 10.1037/0003-066x.54.8.594.

[93] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.

[94] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, *Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers*, 2020. DOI: 10.48550/ARXIV.2002.10957.

[95] J. N. Wulff, G. B. Sajons, G. Pogrebna, S. Lonati, N. Bastardoz, G. C. Banks, *et al.*, "Common methodological mistakes," *The Leadership Quarterly*, vol. 34, no. 1, p. 101 677, Feb. 2023, ISSN: 1048-9843. DOI: 10.1016/j.leaqua.2023.101677.

[96] S. J. Zaccaro, A. L. Rittman, and M. A. Marks, "Team leadership," *The Leadership Quarterly*, vol. 12, no. 4, pp. 451–483, Dec. 2001, ISSN: 1048-9843. DOI: 10.1016/s1048-9843(01)00093-5.

[97] Q. André, "Outlier exclusion procedures must be blind to the researcher's hypothesis.," *Journal of Experimental Psychology: General*, vol. 151, no. 1, pp. 213–223, Jan. 2022, ISSN: 0096-3445. DOI: 10.1037/xge0001069.

[98] J. McCambridge, K. Kypri, and D. Elbourne, "Research participation effects: A skeleton in the methodological cupboard," *Journal of Clinical Epidemiology*, vol. 67, no. 8, pp. 845–849, Aug. 2014, ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2014.03.002.

[99] E. P. Bettinghaus, "Health promotion and the knowledge-attitude-behavior continuum," *Preventive Medicine*, vol. 15, no. 5, pp. 475–491, Sep. 1986, ISSN: 0091-7435. DOI: 10.1016/0091-7435(86)90025-3. [Online]. Available: http://dx.doi.org/10.1016/0091-7435(86)90025-3.

[100] H. Kruger and W. Kearney, "A prototype for assessing information security awareness," *Computers & Security*, vol. 25, no. 4, pp. 289–296, Jun. 2006, ISSN: 0167-4048. DOI: 10.1016/j.cose.2006.02.008. [Online]. Available: http://dx.doi.org/10.1016/j.cose.2006.02.008.

[101] S. Kasten, L. van Osch, M. Candel, and H. de Vries, "The influence of pre-motivational factors on behavior via motivational factors: A test of the i-change model," *BMC Psychology*, vol. 7, no. 1, Feb. 2019, ISSN: 2050-7283. DOI: 10.1186/s40359-019-0283-2.

[102] L. Liu, Y.-P. Liu, J. Wang, L.-W. An, and J.-M. Jiao, "Use of a knowledge-attitude-behaviour education programme for chinese adults undergoing maintenance haemodialysis: Randomized controlled trial," *Journal of International Medical Research*, vol. 44, no. 3, pp. 557–568, Mar. 2016, ISSN: 1473-2300. DOI: 10.1177/0300060515604980.

[103] D. L. Fixsen, S. F. Naoom, K. A. Blase, and R. M. Friedman, *Implementation research: A synthesis of the literature*, National Implementation Research Network, 2005.

[104] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer, "The role of deliberate practice in the acquisition of expert performance.," *Psychological Review*, vol. 100, no. 3, pp. 363–406, Jul. 1993, ISSN: 0033-295X. DOI: 10.1037/0033-295x.100.3.363.

Appendices

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I Zafer Balkan

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Building Task Teams for Incident Response", supervised by Dr. Ricardo Gregorio Lugo
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

18.05.2025

[1]The non-exclusive license is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive license, the non-exclusive license shall not be valid for the period.

Appendix 2 - Teabeleht ja Informeeritud Nõusoleku Vorm

1. Kutsung ja Eesmärk

Sind kutsutakse osalema *simuleeritud küberintsidendi lahendamise lauharjutuses*. Uuringu eesmärgid on kahetised:

- **Äriline** - tugevdada organisatsiooni valmisolekut ja otsustusprotsesse küberkriisi tingimustes.
- **Akadeemiline** - analüüsida grupidünaamikat ja otsustusmehhanisme eelretsenseeritava sotsiaalteadusliku publikatsiooni tarbeks.

Enne otsuse tegemist loe palun alljärgnev teave hoolikalt läbi ja küsi julgelt küsimusi.

2. Mida Osalemine Kaasab?

- 3-tunnine lauharjutus (pluss 15-minutiline paus) **22. aprillil 2025**, kas sessioonil 09:30–12:30 või 14:00–17:00.
- Mängid harjutuse jooksul oma tavapärast operatiiv- või juhtimisrolli, reageerides stsenaariumi inject'idele.
- Sinult palutakse täita isikliku ja meeskonna tasandi intsidiraportid.
  - Meeskonnajuhid koondavad meeskonnatasandi raporti.
- Ühtegi tootmissüsteemi reaalajas ei kasutata.

3. Vabatahtlik Osalemine ja Loobumisõigus

Osalemine on täiesti vabatahtlik. Võid igal ajal ilma põhjendusi esitamata ja tagajärgedeta loobuda. Loobumise korral kustutatakse kõik sinu isikuga seostatavad salvestised ja märkmed, mis pole veel anonüümseks muudetud.

4. Riskid ja Ebamugavused

- Vähene psühholoogiline pinge, mis sarnaneb realistliku ärilise õppusega.
- Võimalik mainekahju, kui individuaalsed tulemused avalikustataks. *Leevendus:* tulemused pseudonümiseeritakse ja esitatakse üksnes koondkujul.

- Kui tunned ebamugavust, võid harjutuse peatada või lõpetada.

5. Kasu

- Kiired organisatsioonilised õppetunnid kübervastupanuvõime parandamiseks.
- Isiklik oskusareng kriisijuhtimises.
- Meeskonna ühise vaimse mudeli ja situatsiooniteadlikkuse tugevdamine.
- Panus akadeemilisse teadmusesse; leiud avalikustatakse anonüümsel kujul.

6. Andmekaitse ja Konfidentsiaalsus

- **Andmekontroller:** <REDACTED>
- **Õiguslik alus:** GDPR art. 6(1)(f) õigustatud huvi & art. 6(1)(e) avalikes huvides tehtav teadustöö; art. 9(2)(j) eriliigiliste andmete töötlemine teaduse eesmärgil.
- **Kogutavad andmed:** intsidiraportid, kirjalikud märkmed, rolliinfo.
- **Säilitamine ja turve:** krüpteeritud hoiustamine EL serverites; juurdepääs ainult uurimisrühmal; säilitusaeg 3 kuud, seejärel kustutamine.
- **Pseudonümiseerimine:** nimed asendatakse koodidega; võtmetabel hoitakse eraldi.
- **Andmeedastus:** andmeid ei edastata väljapoole EL/EER ilma täiendavate kaitsemeetmeteta.

7. Tulemuste Kasutamine ja Võimalik Ärikasutus

Uurimistulemusi võidakse kasutada:

- sisejäreldusraportis organisatsioonile;
- konverentsiettekannetes või teadusartiklites (sotsiaalteadus, küberjulgeolek).

Otseste tsitaatide kasutamiseks küsitakse allpool eraldi nõusolekut.

8. Hüvitis ja Kindlustus

- Rahalist tasu ei maksta; pakutakse kergeid suupisteid ja jooke.
- Tegevus on mitteinvasiivne ja madala riskiga; eraldi tervisekindlustus ei ole vajalik. Kohaldub tööandja vastutuskindlustus.

9. Nõusoleku Kinnitus

Kinnitan, et olen käesoleva teabe läbi lugenud ja aru saanud.

Mõistan, et osalemine on vabatahtlik ja võin igal ajal loobuda.

Annan nõusoleku anonüümsete tsitaatide kasutamiseks akadeemilistes väljaannetes.

Annan nõusoleku minu isikuandmete töötlemiseks ülaltoodud eesmärkidel.

Osaleja nimi: _____ **Kuupäev:** _____

Allkiri: _____

Uurija allkiri: _____ **Kuupäev:** _____

See dokument on koostatud eesti keeles. Ingliskeelne versioon (*"Information Sheet & Informed Consent Form"*) on võrdväärselt kehtiv.

Appendix 3 - Information Sheet & Informed Consent Form

1. Invitation & Purpose

You are invited to take part in a *simulated cyber-incident response exercise*. The purpose is twofold:

- **Business** – to strengthen our organization's preparedness and decision-making under cyber-crisis conditions.
- **Academic** – to analyze group dynamics and decision processes for a peer-reviewed social-science publication.

Before you decide, please read the following information carefully and feel free to ask questions.

2. What Will Participation Involve?

- A 3-hour tabletop session (plus a 15-minute break) facilitated on **22 April 2025**, one of the sessions 09:30–12:30 or 14:00–17:00.
- You will play your normal operational/management role while responding to unfolding scenario injects.
- Incident report documents will be requested to fill in at *person* and *team* level.
- Team leaders are responsible for consolidating a team-level incident report.
- No live production systems will be accessed.

3. Voluntary Participation & Right to Withdraw

Your participation is entirely voluntary. You may withdraw at any point without giving a reason and without penalty. If you withdraw, any recordings or notes that can still be linked to you will be securely deleted.

4. Risks & Discomforts

- Minimal psychological stress comparable to a realistic business exercise.
- Potential reputational concern if individual performance were disclosed. *Mitigation:* results will be pseudonymized and only aggregated in reports.
- If you experience discomfort, you may pause or stop the exercise at any time.

5. Benefits

- Immediate organizational insights to improve cyber-resilience.
- Personal skill development in crisis leadership.
- Building team shared mental model and situational awareness.
- Contribution to academic knowledge; findings will be publicly shared in anonymized form.

6. Data Protection & Confidentiality

- **Data Controller:** REDACTED
- **Legal Basis:** GDPR Art. 6(1)(f) legitimate interest & Art. 6(1)(e) research in the public interest; Art. 9(2)(j) for any special-category data.
- **Data Collected:** Incident reports, written notes, role information.
- **Storage & Security:** Encrypted storage on EU servers; access limited to research team; retention for 3 months, then deletion.
- **Pseudonymisation:** A coded ID key will replace names; the key is stored separately.
- **Data Transfers:** No transfers outside the EU/EEA without additional safeguards.

7. Use of Results & Possible Commercial Uses

Research outputs may be:

- Internal after-action report for the organization.
- Conference papers / journal articles (social-science, cybersecurity).

Separate consent is sought below for direct quotations.

8. Compensation & Insurance

- No monetary compensation is offered; refreshments will be provided.
- The activity is non-interventional and low-risk; separate health insurance is not deemed necessary. Standard employer liability insurance applies.

9. Consent Statement

I confirm that I have read and understood the information above.

I understand participation is voluntary and I may withdraw at any time.

I consent to the use of anonymized quotations in academic publications.

I consent to the processing of my personal data for the purposes outlined.

Participant's Name: _____ **Date:** _____

Signature: _____

Researcher's Signature: _____ **Date:** _____

This document is provided in English. An Estonian translation (*"Teabeleht ja Informeeritud Nõusoleku Vorm"*) will be supplied upon request. Both language versions are of equal validity.

Appendix 4 - Organizational Consent Form for Academic Use

1. Parties

Organization: _____

(*legal name & registration number*)

Authorized signatory (manager): _____

Title/Position: _____

Research team: _____

Institution (University / Research body): _____

Contact (email / phone): _____

2. Description of the Exercise & Study

A cyber-incident response tabletop exercise will be conducted on (date) at (location). The session is designed to:

- Enhance the organization's operational readiness and crisis-management capability.
- Generate anonymized empirical data for a social-science research project analyzing decision-making, teamwork, and organizational learning in cyber-crisis contexts.

3. Scope of Consent

By signing this form, the organization grants the research team a non-exclusive, royalty-free license to collect, analyse, and publish the materials generated during the exercise, including:

- Observation notes, anonymized transcripts, or chat logs.
- Aggregated metrics (e.g., response timelines, decision points).
- De-identified artifacts (playbooks, whiteboard photos, etc.).

Publication outlets may include peer-reviewed journals, academic conferences, teaching materials, and open-access repositories. No confidential business information or personal data identifying the

organization or individual employees will be disclosed without additional written permission.

4. Data Handling & Confidentiality

- Raw data containing business-sensitive details will be stored on an encrypted, access-controlled research drive accessible only to the named researchers.
- De-identification follows GDPR Articles 5(1)(b & c) and relevant institutional data-management policies.
- The organisation may review redacted datasets or request removal of proprietary details prior to publication.

5. Intellectual Property & Commercialization

- Copyright in resulting academic works remains with the authors or their institutions.
- This consent does not transfer any rights to the organization's trademarks, trade secrets, software, or other proprietary assets.
- No commercial use of proprietary information will occur without a separate, explicit agreement.

6. Voluntary Nature & Right to Withdraw

Granting consent is voluntary. The organization may withdraw consent any time **before first publication** by written notice to the researcher. Materials already published prior to withdrawal cannot be retracted, but no further use will be made.

7. Term of Consent

This consent remains valid for three (3) years from the date of signature unless revoked earlier under Section 6.

8. Signatures

Place & Date: _____

Authorized Signatory (signature): _____

Name (print): _____

Title / Position: _____

Researcher (signature): _____

Name & Institution (print): _____

Please retain a copy of this form for your records.

Appendix 5 - Detailed statistical analysis of SRQ1

Question 1a There was no significant main effect of time, $F(1, 14) = 0.176$, $p = 0.681$, $\omega^2 = 0.000$, suggesting that overall scores did not change significantly from pre- to post-training. The interaction between time and group was also not statistically significant, $F(1, 14) = 2.547$, $p = 0.133$, $\omega^2 = 0.015$, although it showed a small effect. These results are reported in Table 12.

Table 12. Within-Subjects Effects — Question 1a

Effect	**df**	**F**	**p**	ω^2
Time	1, 14	0.176	0.681	0.000
Time × Team	1, 14	2.547	0.133	0.015

The between-subjects analysis (Table 13) indicated that overall group membership (control vs. experimental) was not associated with significant differences in responses, $F(1, 14) = 1.217$, $p = 0.289$, $\omega^2 = 0.007$.

Table 13. Between-Subjects Effects — Question 1a

Effect	**df**	**F**	**p**	ω^2
Team	1, 14	1.217	0.289	0.007

Descriptive statistics presented in Table 14 showed a small improvement in the control group's mean score (from $M = 5.222$ to $M = 5.556$), while the experimental group's mean decreased (from $M = 6.429$ to $M = 5.857$). The post hoc test (Table 15) revealed a mean difference of -0.754 between groups (Control - Experimental), which was not statistically significant, $t(14) = -1.103$, $p = 0.289$, with a moderate effect size ($d = -0.514$).

Table 14. Descriptive Statistics — Question 1a

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	5.222	1.856	0.619
Control	Post	5.556	1.667	0.556
Experimental	Pre	6.429	0.787	0.297
Experimental	Post	5.857	1.069	0.404

While not statistically significant, this suggests a moderate trend favoring the experimental group's initial advantage eroding post-training, and a mild gain in control. Although the results were not

statistically significant, the data hint at differential trends between groups worth monitoring in future sessions or with larger sample sizes.

Table 15. Post Hoc Test — Question 1a

Comparison	**Mean Difference**	**t (df = 14)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	-0.754	-1.103	0.289	-0.514

Question 1b There was no statistically significant main effect of time, $F(1, 15) = 1.986$, $p = 0.179$, $\omega^2 = 0.054$, although a small effect size was observed, indicating some variability in responses over time, as shown in Table 16. The interaction between time and group approached significance, $F(1, 15) = 4.353$, $p = 0.055$, with a moderate effect size ($\omega^2 = 0.151$), suggesting a potentially meaningful differential change between the control and experimental groups over time.

Table 16. Within-Subjects Effects — Question 1b

Effect	**df**	**F**	**p**	ω^2
Time	1, 15	1.986	0.179	0.054
Time × Team	1, 15	4.353	0.055	0.151

The between-subjects analysis did not show a statistically significant difference between the control and experimental groups, $F(1, 15) = 0.514$, $p = 0.484$, with essentially no effect size ($\omega^2 = 0.000$), as shown in Table 17.

Table 17. Between-Subjects Effects — Question 1b

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.514	0.484	0.000

Descriptive statistics in Table 18 show a clear increase in the experimental group's mean score from $M = 5.429$ (SD = 0.787) to $M = 6.286$ (SD = 0.951), whereas the control group experienced a slight decline. This positive trend for the experimental group is visually confirmed in Figure 8, which highlights the diverging trajectories of the two groups.

Table 18. Descriptive Statistics — Question 1b

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	5.667	1.000	0.354
Control	Post	5.444	1.014	0.359
Experimental	Pre	5.429	0.787	0.297
Experimental	Post	6.286	0.951	0.359

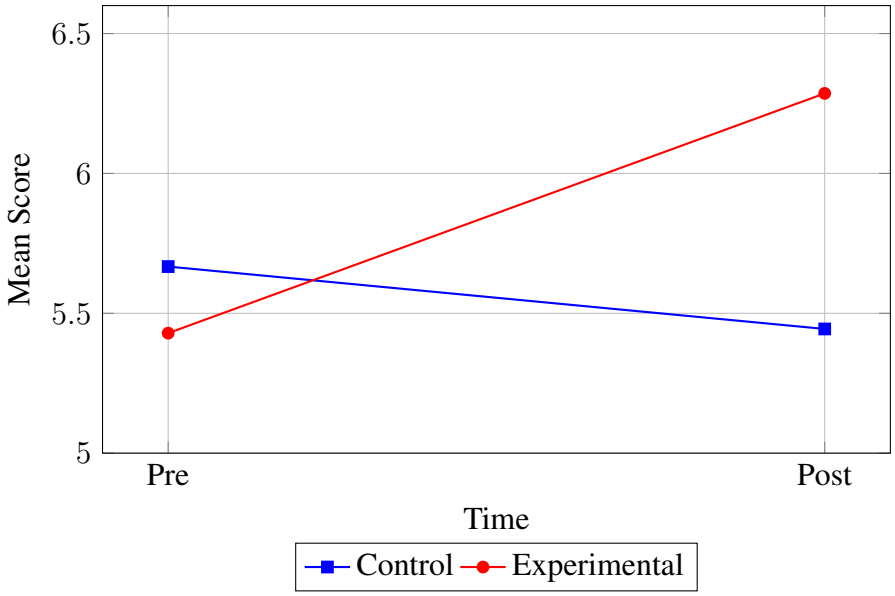

Figure 8. Mean pre- and post-training scores for control and experimental groups on Question 1b.

Table 19. Post Hoc Test — Question 1b

Comparison	**M Diff**	**t (df = 15)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	-0.762	0.717	0.484	0.644

The post hoc test (Table 19) comparing overall means between groups revealed a moderate-to-large effect size ($d = 0.644$), though the difference was not statistically significant, $t(15) = 0.717$, $p = 0.484$. This suggests that while the experimental group's performance improved more noticeably, the variation was not strong enough to reach significance in this sample.

Question 1c There was no significant main effect of time, $F(1, 15) = 0.082$, $p = 0.778$, with a negligible effect size ($\omega^2 = 0.000$), indicating that scores did not differ overall from pre- to post-training. Additionally, there was no significant interaction between time and group, $F(1, 15) = 0.111$, $p = 0.743$, $\omega^2 = 0.000$ (Table 20), suggesting that the pattern of change over time was similar for both groups.

Table 20. Within-Subjects Effects — Question 1c

Effect	**df**	**F**	**p**	ω^2
Time	1, 15	0.082	0.778	0.000
Time × Team	1, 15	0.111	0.743	0.000

The between-subjects analysis revealed no statistically significant group difference, $F(1, 15) = 0.018$, $p = 0.895$, and a near-zero effect size ($\omega^2 = 0.000$), indicating that overall responses were similar across control and experimental teams (Table 21).

Table 21. Between-Subjects Effects — Question 1c

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.018	0.895	0.000

Descriptive statistics (Table 22) show a mild improvement in the experimental group from $M = 6.143$ to $M = 6.286$, while the control group declined slightly from $M = 5.889$ to $M = 5.556$. While not statistically significant, this small improvement in the experimental group is visually illustrated in Figure 9, suggesting a potential positive trend that may warrant monitoring in larger samples.

Table 22. Descriptive Statistics — Question 1c

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	5.889	1.269	0.423
Control	Post	5.556	0.726	0.257
Experimental	Pre	6.143	0.690	0.261
Experimental	Post	6.286	0.951	0.359

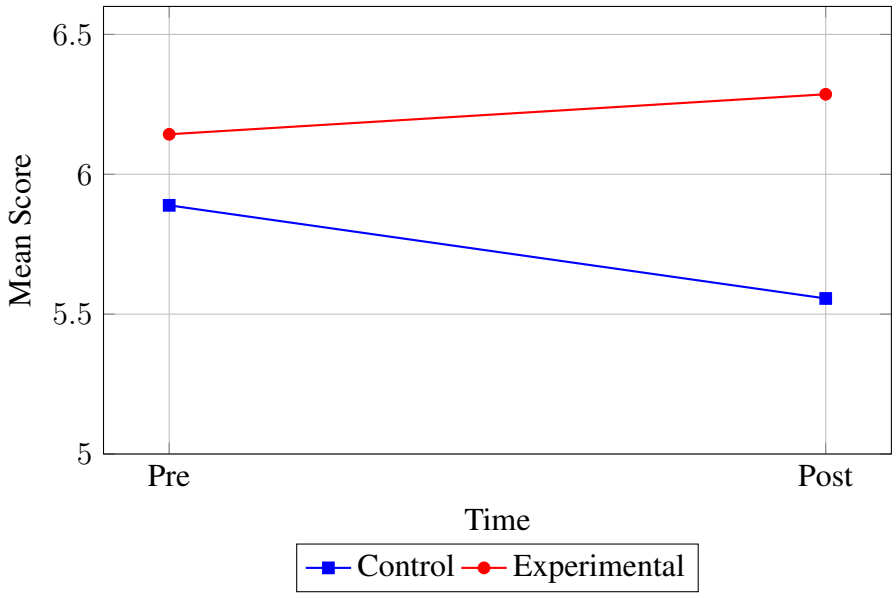

Figure 9. Mean pre- and post-training scores for control and experimental groups on Question 1c.

The post hoc test (Table 23) revealed a negligible difference between groups, $t(15) = 0.133$,

$p = 0.895$, and a very small effect size ($d = 0.106$), supporting the conclusion that group-level differences on this item were minimal in this sample.

Table 23. Post Hoc Test — Question 1c

Comparison	**M Diff**	**t (df = 15)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	-0.143	0.133	0.895	0.106

Question 1d A significant main effect of time was observed, $F(1, 15) = 16.140$, $p = 0.001$, with a large effect size ($\omega^2 = 0.482$), indicating that participants' scores declined significantly from pre- to post-training (Table 24). However, the Time $\times$ Team interaction was not significant, $F(1, 15) = 0.377$, $p = 0.549$, $\omega^2 = 0.000$, suggesting the decline was consistent across both groups.

Table 24. Within-Subjects Effects — Question 1d

Effect	**df**	**F**	**p**	ω^2
Time	1, 15	16.140	0.001	0.482
Time $\times$ Team	1, 15	0.377	0.549	0.000

The between-subjects analysis (Table 25) showed no significant group effect, $F(1, 15) = 0.805$, $p = 0.384$, $\omega^2 = 0.000$, indicating that the overall performance did not differ significantly between control and experimental groups.

Table 25. Between-Subjects Effects — Question 1d

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.805	0.384	0.000

Descriptive statistics (Table 26) showed that the control group's mean score dropped substantially (from $M = 6.444$ to $M = 4.889$), while the experimental group showed a smaller decline (from $M = 6.143$ to $M = 5.571$). This general downward trend is illustrated in Figure 10, where both lines slope downward, but more sharply for the control group.

Table 26. Descriptive Statistics — Question 1d

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	6.444	0.726	0.257
Control	Post	4.889	1.054	0.373
Experimental	Pre	6.143	0.690	0.261
Experimental	Post	5.571	0.787	0.297

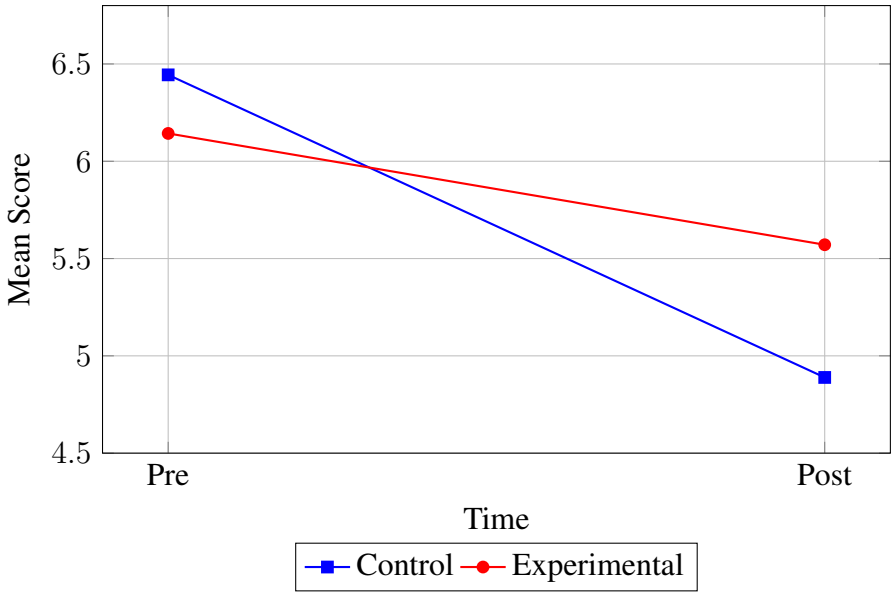

Figure 10. Mean pre- and post-training scores for control and experimental groups on Question 1d.

Table 27. Post Hoc Test — Question 1d

Comparison	**M Diff**	**t (df = 15)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	-0.476	-0.897	0.384	-0.463

The post hoc test (Table 27) showed a non-significant difference between groups, $t(15) = -0.897$, $p = 0.384$, with a moderate effect size ($d = -0.463$). Although the difference is not statistically reliable, the sharper decline in the control group may indicate vulnerability that merits further observation in future iterations.

Question 1e A statistically significant main effect of time was found, $F(1, 15) = 15.130$, $p = 0.001$, with a large effect size ($\omega^2 = 0.459$), suggesting a meaningful decline in scores from pre- to post-training (Table 28). However, the interaction between time and team was not statistically significant, $F(1, 15) = 0.004$, $p = 0.949$, $\omega^2 = 0.000$, indicating that this decline occurred similarly across both control and experimental groups.

Table 28. Within-Subjects Effects — Question 1e

Effect	**df**	**F**	**p**	ω^2
Time	1, 15	15.130	0.001	0.459
Time × Team	1, 15	0.004	0.949	0.000

The between-subjects analysis (Table 29) revealed no significant effect of team membership, $F(1, 15) = 0.136$, $p = 0.718$, with a negligible effect size ($\omega^2 = 0.000$), indicating similar overall response levels between groups.

Table 29. Between-Subjects Effects — Question 1e

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.136	0.718	0.000

Descriptive statistics (Table 30) show a decline in both groups: the control group dropped from $M = 6.444$ to $M = 5.000$, while the experimental group decreased from $M = 6.286$ to $M = 5.714$. The visual representation in Figure 11 confirms a downward slope in both groups' scores, consistent with the significant main effect of time.

Table 30. Descriptive Statistics — Question 1e

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	6.444	0.726	0.257
Control	Post	5.000	1.414	0.500
Experimental	Pre	6.286	0.951	0.359
Experimental	Post	5.714	1.113	0.421

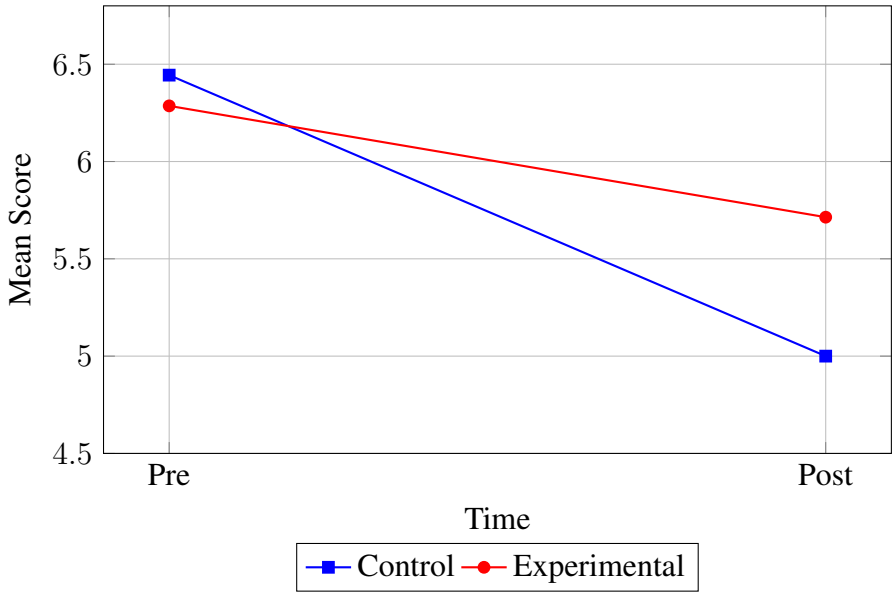

Figure 11. Mean pre- and post-training scores for control and experimental groups on Question 1e.

Table 31. Post Hoc Test — Question 1e

Comparison	**M Diff**	**t (df = 15)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	-0.286	-0.370	0.718	-0.181

The post hoc test (Table 31) confirmed that the difference in scores between groups was not statistically significant, $t(15) = -0.370$, $p = 0.718$, with a small effect size ($d = -0.181$). While both groups declined, the experimental group retained a slight advantage.

Question 2a A highly significant main effect of time was found, $F(1, 14) = 32.810$, $p < 0.001$, with a very large effect size ($\omega^2 = 0.665$), indicating a substantial drop in responses from pre- to post-training (Table 32). The interaction between time and team was not statistically significant, $F(1, 14) = 0.547$, $p = 0.471$, $\omega^2 = 0.000$, suggesting this pattern was similar across both control and experimental groups.

Table 32. Within-Subjects Effects — Question 2a

Effect	**df**	**F**	**p**	ω^2
Time	1, 14	32.810	<.001	0.665
Time × Team	1, 14	0.547	0.471	0.000

Between-subjects analysis (Table 33) showed no significant group effect, $F(1, 14) = 1.143$, $p = 0.304$, with a small effect size ($\omega^2 = 0.016$), indicating that group assignment alone did not meaningfully influence the results.

Table 33. Between-Subjects Effects — Question 2a

Effect	**df**	**F**	**p**	ω^2
Team	1, 14	1.143	0.304	0.016

Descriptive statistics (Table 34) illustrate a sharp decrease in both groups: the control group dropped from $M = 6.444$ to $M = 4.667$, and the experimental group from $M = 6.571$ to $M = 5.286$. This substantial drop, consistent across groups, is visualized in Figure 12.

Table 34. Descriptive Statistics — Question 2a

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	6.444	0.726	0.257
Control	Post	4.667	1.000	0.354
Experimental	Pre	6.571	0.535	0.202
Experimental	Post	5.286	0.951	0.359

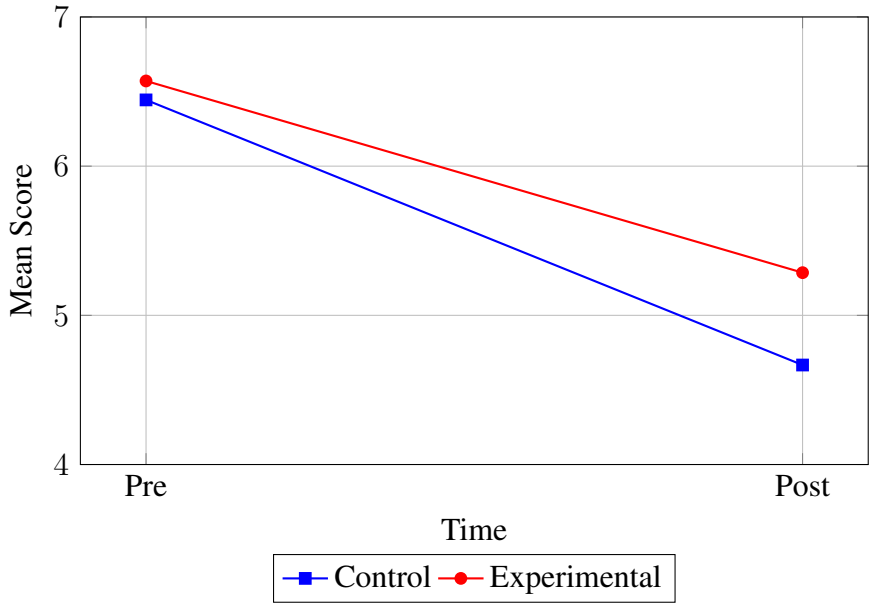

Figure 12. Mean pre- and post-training scores for control and experimental groups on Question 2a.

Table 35. Post Hoc Test — Question 2a

Comparison	**M Diff**	**t (df = 14)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	-0.619	-1.093	0.304	-0.546

Post hoc comparison (Table 35) showed a non-significant difference between groups, $t(14) = -1.093$, $p = 0.304$, with a moderate effect size ($d = -0.546$). While statistical significance was not reached, the direction and size of the effect suggest some differential group dynamics worth further examination.

Question 2b As shown in Table 36, there was a significant main effect of time, $F(1, 15) = 5.137$, $p = 0.039$, with a moderate effect size ($\omega^2 = 0.199$), indicating a meaningful drop in scores from pre- to post-training. The interaction between time and team was not significant, $F(1, 15) = 0.388$, $p = 0.543$, suggesting the pattern was similar across both groups.

Table 36. Within-Subjects Effects — Question 2b

Effect	**df**	**F**	**p**	ω^2
Time	1, 15	5.137	0.039	0.199
Time × Team	1, 15	0.388	0.543	0.000

The between-subjects analysis (Table 37) revealed no significant group differences, $F(1, 15) = 0.123$, $p = 0.730$, $\omega^2 = 0.000$.

Table 37. Between-Subjects Effects — Question 2b

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.123	0.730	0.000

Descriptive statistics in Table 38 show that both groups experienced a drop. The control group declined from $M = 6.000$ to $M = 5.111$, while the experimental group went from $M = 6.143$ to $M = 5.571$. These patterns are visualized in Figure 13.

Table 38. Descriptive Statistics — Question 2b

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	6.000	0.866	0.306
Control	Post	5.111	1.364	0.481
Experimental	Pre	6.143	0.378	0.143
Experimental	Post	5.571	1.272	0.481

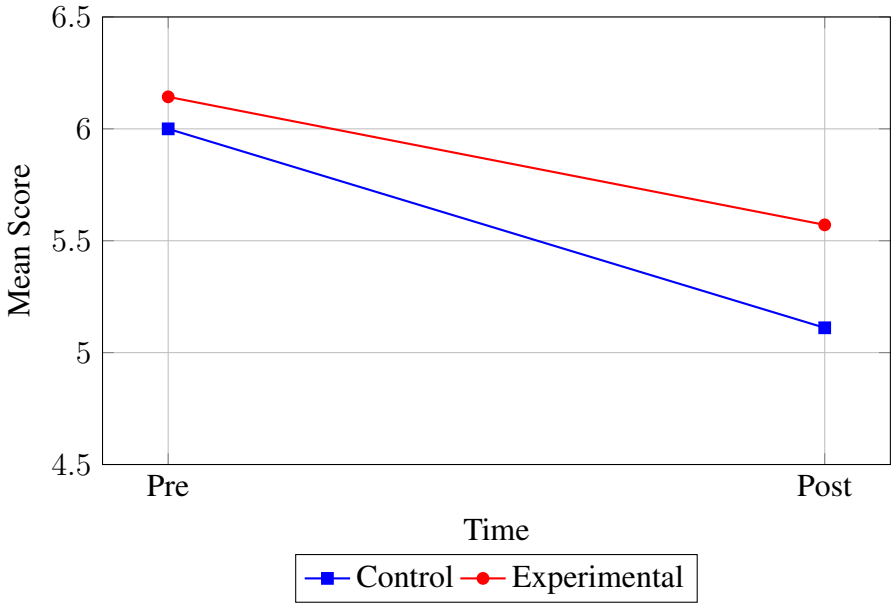

Figure 13. Mean pre- and post-training scores for control and experimental groups on Question 2b.

Table 39. Post Hoc Test — Question 2b

Comparison	**M Diff**	**t (df = 15)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	-0.248	-0.352	0.730	-0.176

Post hoc results (Table 39) showed that the group-level difference was not significant, $t(15) = -0.352$, $p = 0.730$, with a small effect size ($d = -0.176$).

Question 2c As presented in Table 40, the repeated measures ANOVA showed no statistically significant main effect of time, $F(1, 15) = 3.085$, $p = 0.099$, although the effect size was moderate ($\omega^2 = 0.121$), suggesting a possible trend. The interaction between time and team was negligible, $F(1, 15) = 0.045$, $p = 0.834$, with no meaningful effect size ($\omega^2 = 0.000$).

Table 40. Within-Subjects Effects — Question 2c

Effect	**df**	**F**	**p**	ω^2
Time	1, 15	3.085	0.099	0.121
Time × Team	1, 15	0.045	0.834	0.000

Between-subjects differences were not statistically significant either, $F(1, 15) = 0.210$, $p = 0.653$,

as shown in Table 41, and the effect size ($\omega^2 = 0.000$) confirms minimal difference due to group.

Table 41. Between-Subjects Effects — Question 2c

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.210	0.653	0.000

Table 42 and Figure 14 present the descriptive trends. The control group showed a decrease from $M = 6.333$ to $M = 5.889$, while the experimental group improved slightly from $M = 5.857$ to $M = 6.000$. These movements—though small—align with the moderate within-subjects time effect noted earlier, suggesting opposite directional changes.

Table 42. Descriptive Statistics — Question 2c

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	6.333	0.500	0.189
Control	Post	5.889	0.928	0.350
Experimental	Pre	5.857	0.899	0.340
Experimental	Post	6.000	0.816	0.309

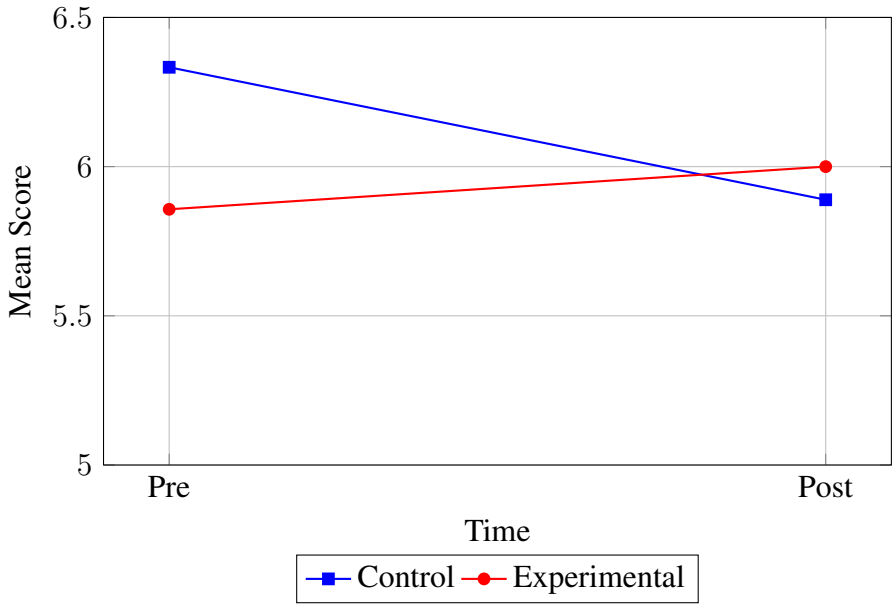

Figure 14. Mean pre- and post-training scores for control and experimental groups on Question 2c.

Table 43. Post Hoc Test — Question 2c

Comparison	**M Diff**	**t (df = 15)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	0.222	0.458	0.653	0.227

Post hoc testing confirmed the lack of group difference, $t(15) = 0.458$, $p = 0.653$, and a small effect size ($d = 0.227$) was observed (Table 43).

Question 6a As shown in Table 44, the repeated measures ANOVA revealed no statistically significant main effect of time, $F(1, 15) = 2.009$, $p = 0.178$, with a small effect size ($\omega^2 = 0.083$). This suggests the training did not result in a reliable overall change in scores for this question. Additionally, the time-by-group interaction was not significant, $F(1, 15) = 0.186$, $p = 0.672$, indicating no differential effect between groups over time.

Table 44. Within-Subjects Effects — Question 6a

Effect	**df**	**F**	**p**	ω^2
Time	1, 15	2.009	0.178	0.083
Time × Team	1, 15	0.186	0.672	0.000

As detailed in Table 45, the between-subjects analysis also showed no significant group-level differences, $F(1, 15) = 0.145$, $p = 0.708$, with a negligible effect size ($\omega^2 = 0.000$).

Table 45. Between-Subjects Effects — Question 6a

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.145	0.708	0.000

Table 46 and Figure 15 show the mean performance scores for each group. The control group showed a decline from $M = 6.000$ to $M = 5.333$, while the experimental group showed a slight decrease from $M = 5.571$ to $M = 5.429$. Although the changes are small, both moved in the same downward direction.

Table 46. Descriptive Statistics — Question 6a

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	6.000	0.816	0.308
Control	Post	5.333	1.000	0.377
Experimental	Pre	5.571	1.134	0.428
Experimental	Post	5.429	0.976	0.369

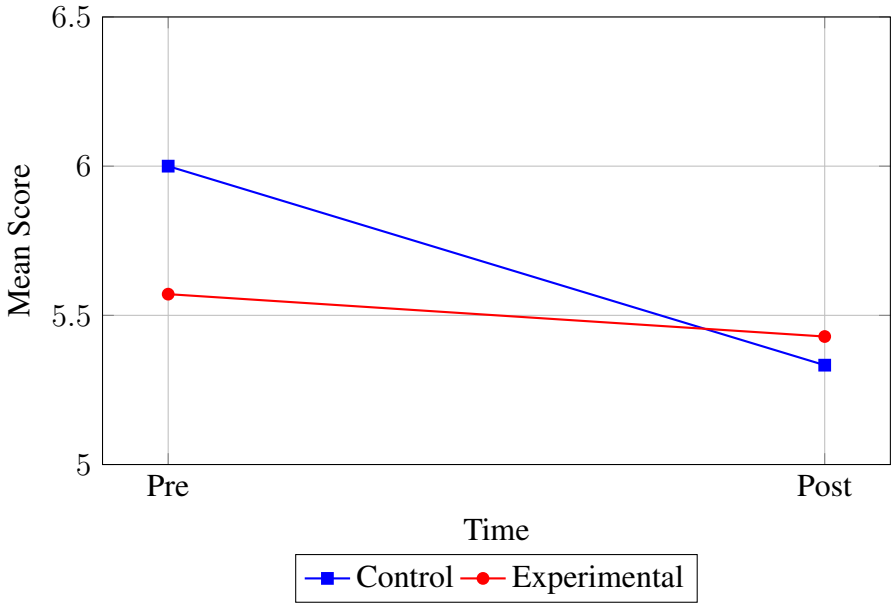

Figure 15. Mean pre- and post-training scores for control and experimental groups on Question 6a.

Table 47. Post Hoc Test — Question 6a

Comparison	M Diff	t (df = 15)	p (Bonf.)	Cohen's d
Control - Experimental	0.190	0.381	0.708	0.194

The post hoc results (see Table 47) suggest that the difference in post-training scores between the control and experimental groups was not significant ($p = 0.708$), and the observed effect size was small ($d = 0.194$). This aligns with earlier ANOVA results, indicating no meaningful group differences or strong training impact for this item.

Question 6b The repeated measures ANOVA revealed no significant main effect of time, $F(1, 15) = 0.226$, $p = 0.641$, with a negligible effect size ($\omega^2 = 0.000$), as shown in Table 48. Likewise, the interaction between time and team was not significant, $F(1, 15) = 0.399$, $p = 0.538$, indicating consistent patterns across both groups.

Table 48. Within-Subjects Effects — Question 6b

Effect	df	F	p	ω^2
Time	1, 15	0.226	0.641	0.000
Time × Team	1, 15	0.399	0.538	0.000

As shown in Table 49, the between-subjects analysis also yielded no significant effect, $F(1, 15) = 0.159$, $p = 0.695$, with no measurable between-group difference ($\omega^2 = 0.000$).

Table 49. Between-Subjects Effects — Question 6b

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.159	0.695	0.000

Descriptive statistics (Table 50) show little change in either group. The control group increased slightly from $M = 5.333$ to $M = 5.556$, while the experimental group remained stable from $M = 5.571$ to $M = 5.429$. These trends are visualized in Figure 16 and support the lack of statistical effects in the ANOVA.

Table 50. Descriptive Statistics — Question 6b

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	5.333	1.225	0.462
Control	Post	5.556	1.667	0.629
Experimental	Pre	5.571	1.134	0.428
Experimental	Post	5.429	0.787	0.297

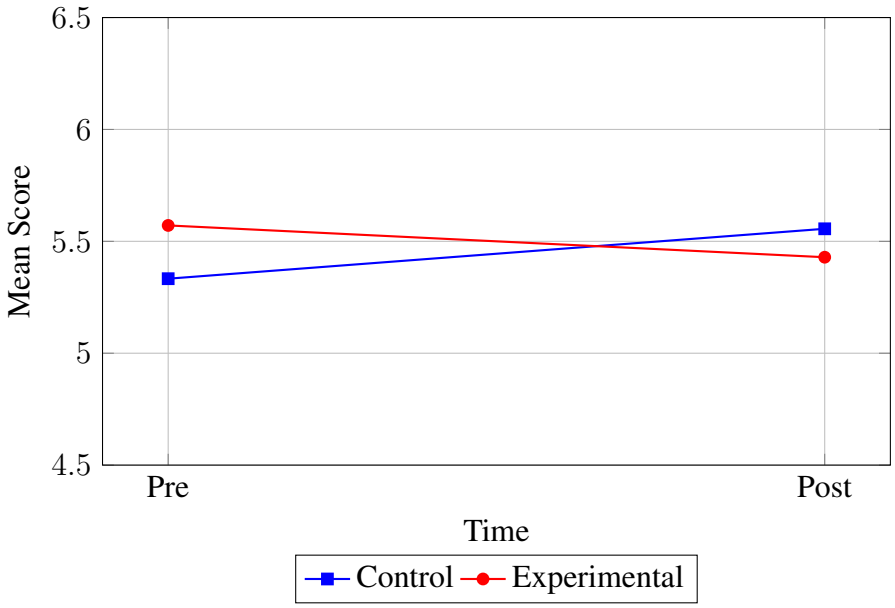

Figure 16. Mean pre- and post-training scores for control and experimental groups on Question 6b.

Table 51. Post Hoc Test — Question 6b

Comparison	**M Diff**	**t (df = 15)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	0.127	0.399	0.695	0.199

The post hoc comparison between groups confirmed the lack of statistical difference, $p = 0.695$, with a very small effect size ($d = 0.199$), reinforcing the earlier conclusion of minimal group separation for this question, as shown in (see Table 51).

Question 6c As shown in Table 52, there was no significant main effect of time, $F(1, 15) = 1.227$, $p = 0.285$, with a small effect size ($\omega^2 = 0.027$), indicating that training did not produce a meaningful overall shift in scores. The time-by-group interaction was also not significant, $F(1, 15) = 0.682$, $p = 0.422$, showing no differential change across teams.

Table 52. Within-Subjects Effects — Question 6c

Effect	**df**	**F**	**p**	ω^2
Time	1, 15	1.227	0.285	0.027
Time × Team	1, 15	0.682	0.422	0.000

Between-subjects effects were also not statistically significant, $F(1, 15) = 0.184$, $p = 0.674$, with negligible group-level differences, as shown in Table 53.

Table 53. Between-Subjects Effects — Question 6c

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.184	0.674	0.000

Table 54 shows that the control group remained steady ($M = 6.000$ both pre- and post-training), while the experimental group showed a small decrease ($M = 6.000$ to $M = 5.429$). Figure 17 illustrates this modest divergence.

Table 54. Descriptive Statistics — Question 6c

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	6.000	0.816	0.308
Control	Post	6.000	0.816	0.308
Experimental	Pre	6.000	0.816	0.308
Experimental	Post	5.429	0.787	0.297

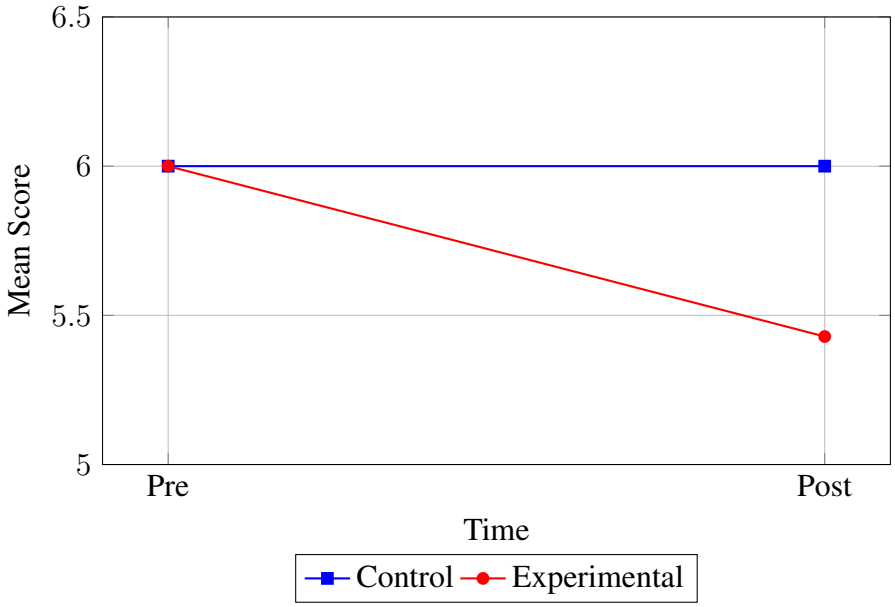

Figure 17. Mean pre- and post-training scores for control and experimental groups on Question 6c.

Table 55. Post Hoc Test — Question 6c

Comparison	**M Diff**	**t (df = 15)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	0.190	0.436	0.674	0.218

The post hoc test confirms no statistically significant difference between the groups ($p = 0.674$), and the effect size was small ($d = 0.218$), consistent with the small changes observed descriptively (see Table 55).

Question 6d As shown in Table 56, there was no significant main effect of time, $F(1, 15) = 0.275$, $p = 0.608$, with a negligible effect size ($\omega^2 = 0.000$). The interaction between time and team was also not statistically significant, $F(1, 15) = 0.040$, $p = 0.844$, indicating no notable differential trends across the groups.

Table 56. Within-Subjects Effects — Question 6d

Effect	**df**	**F**	**p**	ω^2
Time	1, 15	0.275	0.608	0.000
Time × Team	1, 15	0.040	0.844	0.000

Between-subjects effects were not significant either, $F(1, 15) = 0.133$, $p = 0.720$, and the effect size was zero, confirming minimal overall differences between the control and experimental groups (Table 57).

Table 57. Between-Subjects Effects — Question 6d

Effect	**df**	**F**	**p**	ω^2
Team	1, 15	0.133	0.720	0.000

Descriptive statistics (Table 58) show a flat trend: both control and experimental groups had identical pre- and post-training scores ($M = 5.667$ and $M = 5.571$ respectively). These results are illustrated in Figure 18, where both lines remain nearly horizontal.

Table 58. Descriptive Statistics — Question 6d

Group	**Time**	**Mean**	**SD**	**SE**
Control	Pre	5.667	1.000	0.377
Control	Post	5.667	1.000	0.377
Experimental	Pre	5.571	1.134	0.428
Experimental	Post	5.571	1.134	0.428

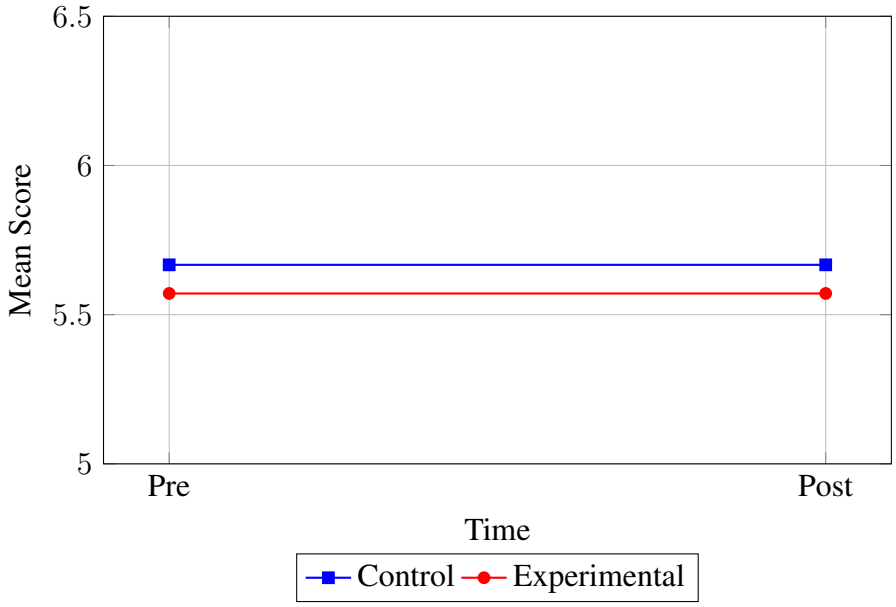

Figure 18. Mean pre- and post-training scores for control and experimental groups on Question 6d.

Table 59. Post Hoc Test — Question 6d

Comparison	**M Diff**	**t (df = 15)**	**p (Bonf.)**	**Cohen's d**
Control - Experimental	0.095	0.366	0.720	0.188

The post hoc comparison in Table 59 confirmed the absence of a meaningful difference between control and experimental groups. The p-value ($p = 0.720$) and the very small effect size ($d = 0.188$) are in line with the flat descriptive and inferential trends observed across the board.

Appendix 6 - Source code for SBERT

The following script automates the extraction of participant-level SBERT cohesion scores from raw free-text responses and prepares the results for downstream analysis in JASP. It processes four input CSV files—each containing `participant_id` and `response` columns corresponding to control and experimental groups, both pre- and post-training. After removing any blank entries, the script employs the *all-MiniLM-L6-v2* sentence transformer model to generate embeddings for each complete response. For each condition, it computes a cosine similarity matrix and averages the pairwise similarities to produce a single cohesion score for each participant at each time point.

Next, the script pivots these long-form scores into a wide table, infers each participant's group (control vs experimental), and computes pre–post differences (*delta_score*). The final output, *participant_scores_wide.csv*, contains exactly the columns needed - *participant_id, group*, *pre_score*, *post_score*, and *delta_score* - for performing paired Wilcoxon and independent Mann–Whitney U tests in JASP.

```
1
2  import pandas as pd
3  import numpy as np
4  from sentence_transformers import SentenceTransformer, util
5
6  # 1. Configuration
7  MODEL_NAME = 'all-MiniLM-L6-v2'
8  dataset = {
9      'ctrl_pre': 'data/control-pre.csv',
10     'ctrl_post': 'data/control-post.csv',
11     'exp_pre': 'data/experimental-pre.csv',
12     'exp_post': 'data/experimental-post.csv'
13 }
14
15 # 2. Initialize SBERT
16 model = SentenceTransformer(MODEL_NAME, device='cpu')
17
18 # 3. Compute per-participant SBERT cohesion scores in "long" form
19 records = []
20 for condition, path in dataset.items():
21     # Load each condition's CSV; expect columns: participant_id,
       response
22     df = pd.read_csv(path, keep_default_na=False)
23     # Drop blank responses
24     valid = df[df['response'].str.strip() != '']
25     if valid.empty:
26         continue
```

```
27
28      pids = valid['participant_id'].tolist()
29      texts = valid['response'].tolist()
30
31      # Embed all valid responses
32      embs = model.encode(
33          texts,
34          convert_to_tensor=True,
35          batch_size=32,
36          show_progress_bar=False
37      )
38
39      # Cosine-similarity matrix
40      sim_mat = util.cos_sim(embs, embs).cpu().numpy()
41      np.fill_diagonal(sim_mat, np.nan)
42
43      # Row-wise mean -> one score per participant
44      scores = np.nanmean(sim_mat, axis=1)
45
46      # Record results
47      for pid, score in zip(pids, scores):
48          records.append({
49              'participant_id': pid,
50              'condition': condition,
51              'score': float(score)
52          })
53
54  # 4. Build a long-form DataFrame
55  df_scores = pd.DataFrame.from_records(records)
56
57  # 5. Pivot to wide form: one column per condition
58  df_wide = df_scores.pivot(index='participant_id', columns='condition',
        values='score')
59
60  # 6. Infer group membership
61  df_wide['group'] = np.where(df_wide['ctrl_pre'].notna(), 'control', '
        experimental')
62
63  # 7. Assemble pre, post, and delta columns
64  df_wide['pre_score'] = df_wide['ctrl_pre'].fillna(df_wide['exp_pre'])
65  df_wide['post_score'] = df_wide['ctrl_post'].fillna(df_wide['exp_post'
        ])
66  df_wide['delta_score'] = df_wide['post_score'] - df_wide['pre_score']
67
68  # 8. Select and reorder for export
69  out = df_wide.reset_index()[[
70      'participant_id',
```

```
    'group',
    'pre_score',
    'post_score',
    'delta_score'
]]

# 9. Save wide CSV for JASP
out.to_csv('participant_scores_wide.csv', index=False)
print(f"Wrote {len(out)} participants to 'participant_scores_wide.csv'"
    )
```

Listing 1. Source code for data preparation running SBERT against free-text raw data

Appendix 7 - Checklists for SRQ2

This section reproduces the gold-standard models derived from the exercise scenarios, together with the checklists used to measure **Team Mental Model (TMM) Accuracy** for SRQ2 based on the content of the team-leader reports.

How to Use the Checklists

- **Review team-leader reports** - read each of the four reports (Team A/Scenario A, Team A/Scenario B, Team B/Scenario A, Team B/Scenario B).
- **Apply the relevant checklist** - use Checklist A for Scenario A reports and Checklist B for Scenario B reports.
- **Score each item**
  - Yes (1 point) - the report clearly and correctly addresses the item.
  - Partial (0.5 points) - the report mentions the item but is inaccurate, incomplete, or vague.
  - No (0 points) - the report omits or fully misstates the item.
- **Calculate the total score** - sum the item scores to obtain an overall accuracy score for each report.

Scenario A: Supply-Chain Attack via VS Code Extension

#	Checklist Item	Answer	Score
1	Correctly identified the Kill Chain phase?	7	
2	Correctly identified the operational impact?	2	
3	Correctly identified event class?	Data breach	
4	Correctly identifies the initial vector as a software update/extension issue?		
5	Mentions the installation or presence of an unexpected/malicious service?		
6	Identifies unusual network traffic/tunnelling (e.g. to `*.dev.tunnels.ms`)?		
7	Mentions discovery/use of credentials found locally (e.g. Notepad++ backup)?		
8	Recognises potential or confirmed data exfiltration?		
9	Recommends isolation of affected systems (dev machine *and* server)?		
10	Acknowledges potential compliance implications (e.g. PCI DSS)?		

Total Score (Scenario A):

Scenario B: Ransomware via Web Shell & VM Escape

#	Checklist Item	Answer	Score
1	Correctly identified the Kill Chain phase?	7	
2	Correctly identified the operational impact?	4	
3	Correctly identified event class?	Ransomware	
4	Correctly identifies the initial vector as a web-shell upload?		
5	Mentions the compromise of the hypervisor itself (VM escape)?		
6	Identifies the loss/compromise of critical infrastructure (e.g. SIEM)?		
7	Explicitly advises **against** immediate shutdown of compromised server to allow live forensics?		
8	Suggests restoring affected VMs from backups as the primary recovery method?		
9	Acknowledges the need to use alternative logging sources due to SIEM compromise?		
10	Mentions patching/rebuilding the compromised hypervisor as part of remediation?		

Total Score (Scenario B):