

Research paper

A Systematic Literature Review of Capture the Flag (CTF) Methodologies in Developing Cybersecurity Competence

Triin Muulmann ^{1*} , Ricardo Gregorio Lugo ^{1,2} , Stefan Sütterlin ^{2,3} 

¹ Tallinn University of Technology, ESTONIA

² Ostfold University College, NORWAY

³ Albstadt-Sigmaringen University, GERMANY

*Corresponding Author: triin.muulmann@taltech.ee

Citation: Muulmann, T., Lugo, R. G., & Sütterlin, S. (2026). A systematic literature review of capture the flag (CTF) methodologies in developing cybersecurity competence. *European Journal of STEM Education*. <https://doi.org/10.20897/ejsteme/19494>

Published: October 6, 2026

ABSTRACT

As cyber threats continue to expand across information technology (IT) and operational technology (OT) domains, cybersecurity education increasingly relies on practice-oriented methods beyond traditional theoretical instruction. Capture the Flag (CTF) competitions have emerged as a widely adopted form of experiential and gamified learning that aims to develop both technical and transversal cybersecurity competences. This systematic literature review searched for empirical studies published between 2015 and 2025 and synthesizes evidence from 18 eligible studies published between 2017 and 2025 examining the use of CTF methodologies in cybersecurity education. The review maps CTF formats (Jeopardy-style, Attack–Defense, hybrid, and immersive), educational contexts, target populations, and reported learning outcomes. It also evaluates the methodological quality and certainty of the available evidence using the GRADE framework. Across the reviewed studies, CTF participation was frequently associated with increased learner engagement and self-reported confidence. Some studies also reported short-term improvements in technical skills and cybersecurity awareness; however, these findings were generally based on self-reported measures, indirect performance indicators, small samples, or uncontrolled study designs. Several recurring design features were identified across the included studies, including scaffolded difficulty, teamwork, and progressive challenges; however, their independent effects on learning were rarely evaluated. Nevertheless, the current evidence base is limited by heterogeneous methodologies, inconsistent outcome measures, and the lack of longitudinal evaluation. Overall, CTF activities were frequently associated with learner engagement and perceived confidence, although the certainty of evidence for these outcomes was Low. Evidence for demonstrated technical competence, learning retention, and transfer to practice was Very Low. More rigorous comparative and longitudinal research is therefore needed to determine the educational effectiveness of CTF-based learning and its potential contribution to sustained cybersecurity competence.

Keywords: capture the flag (CTF), cybersecurity education, virtual reality, gamification, cybersecurity competence, operational technology, experiential learning, IT/OT integration, systematic literature review

Cybersecurity threats have escalated in complexity, spanning both information technology (IT) and operational technology (OT) domains, which necessitates the use of experiential learning approaches beyond traditional theoretical instruction (NIST, 2025; Stouffer et al., 2023). As a result, cybersecurity education is increasingly

expected to develop not only conceptual knowledge but also practical, transferable skills that can be applied in real-world contexts (Ahmed et al., 2024; Alnajim et al., 2023).

CTF-based learning also intersects with a broader family of game-based educational approaches, including serious games, educational escape rooms, simulation games, and scenario-based incident-response exercises (Kalogiannakis et al., 2021; Zouahi, 2023). Although these approaches differ in their degree of competition, collaboration, realism, and technical complexity, they share the common goal of promoting active participation, feedback, and authentic problem-solving (Kalogiannakis et al., 2021; Malone et al., 2021). Systematic reviews of gamification in education suggest that game elements can enhance learner motivation and engagement when they are aligned with clearly defined learning objectives and appropriate instructional design (Kalogiannakis et al., 2021).

Among emerging instructional approaches, Capture-the-Flag (CTF) competitions have gained prominence as a form of gamified, challenge-based learning (Švábenský et al., 2021; Vykopal et al., 2020).

These environments simulate realistic cybersecurity scenarios and require participants to apply technical knowledge, problem-solving skills, and teamwork under time constraints (Vykopal et al., 2020). Jeopardy-style CTFs, which consist of independent challenges across domains such as cryptography and web security, are widely used in higher education and have been reported to support active learning and technical competence (Leune & Petrilli Jr, 2017; Vykopal et al., 2020).

Recent developments have expanded CTF formats into narrative-driven and escape-room-style designs that emphasize critical thinking, ethical reasoning, and collaboration (Zouahi, 2023). These approaches embed challenges within structured scenarios, increasing immersion and contextual understanding compared to traditional task-based formats (Zouahi, 2023). Furthermore, CTF-based learning has been adopted beyond computer science, including engineering and industrial control system (ICS) domains, where it supports interdisciplinary cybersecurity training (Karampidis et al., 2023; Leune & Petrilli Jr, 2017). CTF methodologies can be broadly categorized into three main types: (1) Jeopardy-style CTFs, which consist of independent challenges; (2) Attack–Defense CTFs, where participants simultaneously attack and defend systems in real time; and (3) hybrid or scenario-based formats, which combine structured challenges with narrative or system-level contexts (Švábenský et al., 2021; Vykopal et al., 2020).

In this review, CTF methodologies are treated as the primary pedagogical focus, while related technologies such as cyber ranges, VR/AR, and OT environments are considered as contextual or enabling components rather than core constructs. Within this broader landscape, emerging technologies such as virtual and augmented reality (VR/AR) are being explored to enhance CTF environments by enabling immersive, spatial interaction with simulated systems. While such approaches show promise in improving engagement and experiential learning, empirical evidence regarding their impact on deeper cognitive outcomes remains limited (Alnajim et al., 2023).

Despite the growing adoption of CTF methodologies, the existing body of literature remains fragmented (Ahmed et al., 2024; Demetrio et al., 2019).

Studies differ significantly in terms of design, participant populations, outcome measures, and methodological rigor, which limits comparability and synthesis of findings (Demetrio et al., 2019; V. S. Ford, 2017). Although many studies report improvements in engagement, skills, and awareness, the lack of standardized evaluation frameworks and consistent reporting practices makes it difficult to draw robust conclusions about the effectiveness of CTF-based learning (Ahmed et al., 2024; V. S. Ford, 2017).

While prior studies and reviews have examined aspects of cybersecurity education or gamified learning, few have systematically synthesized CTF methodologies across contexts while simultaneously evaluating the strength and certainty of evidence. As a result, there remains limited understanding of whether different CTF designs are associated with different educational outcomes and under what conditions competence development is assessed. Given this fragmentation, there is a clear need for a systematic review that not only maps the landscape of CTF-based learning but also critically evaluates the quality of evidence and identifies patterns across implementations. Such a synthesis can contribute to both research and practice by identifying recurring design features, reported outcomes, areas where evidence remains limited, and methodological improvements needed in future studies. Despite these developments, the evidence base surrounding CTF-based learning remains uneven.

This study addresses this gap by conducting a systematic literature review of CTF methodologies in cybersecurity education. The review has three primary objectives: (1) to identify and categorize the types and formats of CTF implementations and their application contexts; (2) to synthesize reported learning outcomes and the methods used to measure them; and (3) to assess the methodological quality and certainty of evidence using the GRADE framework.

BACKGROUND

Cybersecurity education and the role of CTFs

CTF methodologies are commonly categorized into three main types: Jeopardy-style CTFs, which consist of independent challenges across domains; Attack–Defense CTFs, where teams simultaneously attack and defend systems; and hybrid formats that combine elements of both (Vykopal et al., 2020);(Švábenský et al., 2021).

In recent years, Capture the Flag (CTF) competitions have become a cornerstone of cybersecurity education, offering experiential, challenge-driven environments that extend beyond traditional pedagogical approaches. CTFs immerse participants in practical exercises—ranging from web exploitation and password cracking to network breaches—that require them to apply technical knowledge in a dynamic, gamified setting (Vykopal et al., 2020). As cybersecurity threats grow in scope and sophistication, education systems face increasing pressure to equip learners with the skills necessary to detect, respond to, and mitigate cyber incidents. One increasingly used approach in this domain is the CTF competition format, which provides hands-on, challenge-based environments for practising technical and strategic cybersecurity skills (Vykopal et al., 2020). CTFs simulate real-world scenarios and require participants to solve tasks in areas such as cryptography, binary exploitation, forensics, and web security (Leune & Petrilli Jr, 2017). These exercises promote experiential learning and problem-solving through practice rather than passive instruction, providing opportunities for practical application by learners at different levels of experience (Lazarov et al., 2025).

Cyber ranges and CTF platforms provide controlled, isolated environments where learners can engage in cybersecurity exercises without risking production systems (Organisation., 2020). Cyber ranges replicate IT/OT infrastructures for scenario-based, team-oriented training, whereas CTF platforms deliver gamified, challenge-driven tasks (flags, scoring, leaderboards); CTFs can run on top of a cyber range, but they are not synonymous (Organisation., 2020). Research shows that well-designed gamification elements, such as points, leaderboards, and narrative missions, can boost engagement and motivation, though effects vary by learner and design (NIST., 2025; Stouffer et al., 2023)

Integrating operational technology (OT) in cybersecurity training

The convergence of information technology (IT) and operational technology (OT) has created new educational requirements that extend beyond traditional IT-focused cybersecurity curricula (Stouffer et al., 2023); (NIST., 2025). OT environments operate critical infrastructure such as energy, manufacturing, and water systems, where cybersecurity incidents may directly affect physical processes and safety (Stouffer et al., 2023).

Consequently, cybersecurity education increasingly incorporates OT concepts and cyber-physical systems into practical training (NIST., 2025; Stouffer et al., 2023). Within this context, OT-aware cyber ranges and Capture-the-Flag (CTF) environments provide safe and controlled settings in which learners can practice incident detection, response, and recovery without affecting operational infrastructure (Enisa., 2018; Stouffer et al., 2023).

Although only a small number of studies included in this review focused specifically on OT or industrial control systems (ICS), these studies suggest that realistic system-level scenarios may strengthen contextual understanding and situational awareness (Antonioli et al., 2017; Karampidis et al., 2023; Korkiakoski et al., 2023; Lazarov et al., 2025). However, further empirical research is needed to determine whether OT-oriented CTF environments provide measurable educational advantages compared with conventional CTF formats (Ahmed et al., 2024; Alnajim et al., 2023; Karampidis et al., 2023).

Competence development through CTF participation

Cybersecurity competence involves the integration of knowledge, technical skills, problem-solving abilities, and the capacity to apply these capabilities in authentic or simulated cybersecurity contexts (Gervais, 2016; Mulder, 2014). Capture-the-Flag activities provide opportunities for learners to engage with practical cybersecurity tasks, but successful task completion does not necessarily demonstrate broader or transferable cybersecurity competence. Performance indicators such as flags captured, task-completion rates, or time-to-solve may reflect task-specific performance and should therefore be distinguished from validated measures of knowledge, skills, or competence.

A further distinction is required between perceived and demonstrated learning. Several CTF studies assess outcomes through self-reported confidence, self-efficacy, motivation, or perceived knowledge, whereas others use task performance, knowledge assessments, or behavioural indicators. These measures represent different constructs and should not be interpreted as interchangeable. In particular, increases in perceived confidence or self-efficacy do not necessarily indicate corresponding improvements in objectively demonstrated cybersecurity competence. This distinction is therefore important when interpreting learning outcomes across heterogeneous CTF studies.

Learning theories such as self-efficacy theory (Bandura & Wessels, 1997), sociocultural perspectives on collaborative learning (Vygotsky, 1987), and Cognitive Load Theory (Sweller et al., 2019) provide possible interpretive frameworks for understanding features commonly observed in CTF environments, including progressive challenges, teamwork, feedback, and scaffolding. However, these theoretical mechanisms were rarely tested directly in the studies included in this review. They are therefore used here as interpretive perspectives rather than as empirically established explanations for the educational effects of CTF participation.

Accordingly, this review distinguishes between perceived outcomes, demonstrated performance, and evidence of longer-term competence development. This distinction informs the synthesis of learning outcomes and the interpretation of evidence concerning engagement, confidence/self-efficacy, technical competence, retention, and transfer to practice.

METHODOLOGY

Research design

This study employs systematic literature review methodology to synthesize empirical research on the use of Capture-the-Flag (CTF) methodologies in cybersecurity education. Recent systematic reviews in educational technology have emphasized the importance of transparent reporting, reproducible search strategies, and clearly defined eligibility criteria when synthesizing rapidly evolving technology-supported learning research (İpek et al., 2023; Page et al., 2021). These methodological principles are particularly relevant in cybersecurity education, where educational technologies, instructional approaches, and evaluation methods continue to evolve rapidly (Ahmed et al., 2024; Alnajim et al., 2023). Accordingly, this review follows the PRISMA 2020 reporting guidelines to enhance transparency, reproducibility, and methodological rigor.

Research questions

This systematic review is guided by the following research questions:

RQ1: How are Capture-the-Flag (CTF) methodologies designed and implemented in cybersecurity education, in terms of formats, technological features, and application contexts?

RQ2: What learning outcomes are associated with CTF-based cybersecurity education, and how are these outcomes operationalized and measured across studies?

RQ3: What is the methodological quality and strength of evidence in studies evaluating CTF-based learning, and what limitations affect the reliability of reported outcomes?

Together, these questions enable a structured synthesis of both the design characteristics and the evidential strength of CTF-based learning in cybersecurity education.

Eligibility criteria

Studies were eligible if they reported empirical findings on the design, implementation, or evaluation of Capture-the-Flag (CTF) methodologies in cybersecurity education or training. Eligible study designs included intervention studies, quasi-experimental studies, observational evaluations, design-based research with empirical evaluation, and field implementations that reported educational outcomes, usability findings, or learner performance. Studies describing CTF platforms without empirical evaluation, conceptual papers, opinion articles, tutorials, editorials, and purely technical system descriptions were excluded. Although the review primarily focused on learning outcomes, design-oriented studies were retained where they contributed evidence regarding educational implementation or learner interaction with CTF environments.

For the purposes of this review, empirical educational evidence was defined as systematically collected data relevant to at least one research question, including participant learning outcomes, task or performance data, usability data, interaction logs, engagement indicators, implementation observations, or expert evaluation of an educational implementation. Papers describing only the technical architecture or potential educational use of a platform without reporting such empirical evidence were excluded.

Search strategy

A systematic literature search was conducted in three multidisciplinary bibliographic sources: Google Scholar, Scopus, and Web of Science. Searches were initially performed on 16 July 2025 and updated on 28 November 2025 to identify newly published studies. The search strategy was developed around four core concepts: Capture the Flag (CTF), cybersecurity, education/training, and learning.

Because each database uses different indexing systems, field codes, and search syntax, database-specific search strings were developed rather than applying an identical Boolean expression across all sources. Google Scholar was searched using title and full-text searches, Scopus using the TITLE-ABS-KEY field, and Web of Science using the

Topic Search (TS) field. The complete search syntax for each database is provided in Supplementary Table S1 available online at: <https://osf.io/ax24y/files/ahsb4>.

The Google Scholar searches yielded a total of 250 records, all of which were screened. No additional numerical screening limit was applied. Searches retrieved 250 Google Scholar records, 37 Scopus records, and 119 Web of Science records. Database-specific search strategies reflected differences in database indexing, search operators, and field availability rather than differences in review scope.

All records were exported into a master spreadsheet. Duplicate records were identified using combinations of title, authors, publication year, DOI, and publication venue. Potential duplicates were manually verified before removal to minimise erroneous exclusions. Each remaining record received a unique screening identifier before title and abstract screening commenced.

Design-oriented empirical studies, including platform implementation reports and field evaluations, were retained when they provided empirical educational evidence despite not always representing conventional peer-reviewed intervention studies.

Study selection

Following duplicate removal, all remaining records underwent title and abstract screening according to predefined eligibility criteria. A standardized Excel spreadsheet was used throughout the screening process to document bibliographic information, publication characteristics, and reviewer decisions for each retrieved study.

Studies were considered eligible if they were published in English between 2015 and 2025, the full text was available in an electronic format, and they were published as peer-reviewed journal articles or full conference papers reporting empirical findings related to Capture-the-Flag (CTF) methodologies in cybersecurity education or training. Publications such as books, book chapters, theses, dissertations, technical reports, manuals, tutorials, editorials, opinion papers, consultancy reports, preprints, and other non-peer-reviewed literature were excluded. Studies that did not include an empirical evaluation or were unrelated to CTF-based cybersecurity education were also excluded.

The first reviewer screened all 386 records remaining after duplicate removal. Records classified as potentially eligible or uncertain were independently verified by the second reviewer; however, the exact number of records subjected to this verification was not recorded separately and therefore cannot be reported retrospectively. Following this stage, 136 reports were sought for retrieval, of which 98 full-text reports were assessed for eligibility.

During title and abstract screening, records were classified using a traffic-light coding approach. Records that clearly met the inclusion criteria were classified as potentially eligible, records that failed one or more inclusion criteria were excluded, and records for which eligibility remained uncertain were identified for further review.

The first reviewer conducted the full-text eligibility assessment, and the second reviewer verified the eligibility decisions for the full-text records. The second reviewer had access to the first reviewer's screening decisions when conducting the verification. Any disagreements regarding eligibility were resolved through discussion and consensus. The number of disagreements was not recorded systematically; therefore, inter-reviewer agreement could not be quantified. Studies fulfilling all eligibility criteria were subsequently included in the qualitative synthesis. The complete study-selection process is illustrated in the PRISMA flow diagram.

Data extraction

A standardized Excel-based data extraction form was developed before the screening process to ensure consistency across all included studies. During the initial screening, bibliographic information including publication source, publication type, authors, publication year, article title, journal or conference name, DOI, keywords, abstract, and APA reference was recorded for each potentially eligible study.

Following study selection, additional methodological and educational characteristics were extracted from the included studies. These included participant population, educational setting, CTF format, application domain, outcome measures, assessment instruments, implementation characteristics (e.g., scaffolding, hints, teamwork, and technological features), follow-up period, and reported learning outcomes.

To improve reliability, the first reviewer completed the initial data extraction for all included studies. The second reviewer independently verified the extracted information, and any discrepancies were discussed until consensus was reached before the final synthesis.

Effect measures

Where available, quantitative data were extracted in the form of mean differences (pre-post or between-group comparisons), standardized mean differences to account for scale heterogeneity, proportional measures such as task success rates, and time-based performance indicators (e.g., time-to-first-flag). Qualitative findings were

synthesized using a thematic approach, with themes systematically mapped to the predefined outcome framework to ensure coherence between quantitative and qualitative evidence.

Synthesis methods

Given the diversity of study designs and outcome measures, a narrative synthesis was undertaken, stratified by CTF format and educational or professional context (higher education, K–12 and outreach, OT/ICS training, and industry). Where three or more studies reported comparable quantitative metrics, results were summarized in terms of central tendency and dispersion to facilitate cross-study comparison. A meta-analysis was not conducted, as substantial heterogeneity in outcomes, measurement instruments, and reporting formats precluded meaningful statistical aggregation.

Study-level quality assessment

Because the included evidence comprised heterogeneous study designs—including educational interventions, observational studies, mixed-method evaluations, pilot studies, quasi-experimental studies, experience reports, and design-oriented empirical studies—the first author developed a structured review-specific appraisal framework for this review. The framework was developed before the final study-level appraisal and was designed to enable a consistent descriptive comparison across the heterogeneous evidence base. It was not formally piloted or validated as an independent risk-of-bias instrument.

Established critical appraisal approaches were considered when developing the assessment framework. However, available instruments are generally tailored to methodological designs or require different criteria for different study types.

For example, JBI provides design-specific critical appraisal tools, including a dedicated tool for quasi-experimental studies (Barker et al., 2024), while the Mixed Methods Appraisal Tool (MMAT) accommodates qualitative, quantitative, and mixed-method study designs (Hong et al., 2018).

Given that the present review also included educational field implementations, pilot studies, experience reports, and design-oriented empirical evaluations, no single established instrument was considered sufficiently applicable across the complete evidence base. A review-specific framework was therefore used to provide a common descriptive assessment rather than to replace design-specific risk-of-bias instruments.

The appraisal focused on four domains that could be applied across the included study designs: participant selection, outcome measurement, reporting transparency, and protection against bias. The criteria were interpreted in relation to study design. For example, the absence of randomisation was not treated identically in an observational study and an educational intervention, but the implications of uncontrolled designs, confounding, selection processes, and measurement limitations were reflected in the protection-against-bias judgement.

The first author conducted the initial appraisal using the predefined framework. The second reviewer subsequently verified all ratings with access to the initial assessments, and discrepancies were discussed and resolved through consensus. Because the framework was not formally piloted or independently validated, the resulting ratings are presented as structured review-specific methodological judgements rather than objective or validated risk-of-bias estimates see [Table 1](#).

Table 1

Quality appraisal rubric

Domain	High	Moderate	Low
Selection	Clearly defined sampling strategy with appropriate participants	Some selection limitations or incomplete reporting	Convenience sample or poorly described selection
Measurement	Validated or objective outcome measures	Mixed validated and non-validated measures	Mainly subjective or non-validated measures
Reporting	Complete description of methods and results	Minor omissions	Important methodological details missing
Protection against bias	Design and procedures provide relatively strong protection against major sources of bias	Some important threats to validity or confounding remain	Major threats from uncontrolled design, confounding, selection, measurement, or other sources of bias

Note. For all four domains, High represents the most favorable methodological assessment and Low the least favorable. For the “Protection against bias” domain specifically, High indicates relatively strong protection against major sources of bias, whereas Low indicates substantial vulnerability to bias. Overall methodological quality was determined using the decision rule described in Section 3.9 and was not calculated as a numerical average of domain ratings.

Overall methodological quality was classified as High, Moderate, or Low. A High overall rating required predominantly High domain ratings, with no Low rating for measurement or protection against bias. A Moderate rating was assigned when the study showed a mixture of High and Moderate ratings or contained a single Low rating that did not substantially undermine interpretation of the educational outcomes. A Low overall rating was assigned when the study had a Low rating for measurement or protection against bias that substantially limited interpretation, or when two or more methodological domains were rated Low. Overall ratings were therefore based on the pattern and importance of methodological limitations rather than a simple numerical majority rule.

Assessment of certainty of evidence

Certainty of evidence was assessed at the outcome level using the GRADE framework (Grading of Recommendations Assessment, Development and Evaluation), considering risk of bias, inconsistency, indirectness, imprecision, and publication bias where assessable (GRADE., 2025). Because the evidence base consisted predominantly of non-randomized educational studies, certainty generally began at Low. No outcome domain met the criteria for rating up the certainty of evidence. Certainty was downgraded where substantial concerns were identified across the GRADE domains.

Before the assessment, findings were grouped into five outcome domains according to the nature and timing of the reported outcomes. Engagement and motivation included participation indicators, task engagement, motivation ratings, submission activity, and related learner-experience measures. Confidence/self-efficacy included perceived confidence, perceived capability, and explicitly measured self-efficacy. These constructs were grouped for synthesis because of the limited evidence base but were not considered theoretically equivalent.

Knowledge and technical competence included objective or performance-based indicators intended to represent cybersecurity knowledge or skill. Platform metrics such as flags submitted, task completion, solve rates, or leaderboard performance were treated as indirect indicators unless linked to an explicit assessment of competence. Retention referred to learning outcomes assessed after a delay following the intervention rather than immediately after training. Transfer to practice referred to evidence that learning acquired through CTF participation was applied to a different task, authentic professional setting, workplace context, or another situation beyond the original training activity.

Given the heterogeneity of study designs and measurement instruments, inconsistency was assessed qualitatively by examining the direction and coherence of findings across studies rather than through statistical heterogeneity estimates. Publication bias could not be formally evaluated because of the small number and methodological diversity of studies contributing to each outcome domain. Final certainty ratings were reported as High, Moderate, Low, or Very Low. In the final assessment, engagement/motivation and confidence/self-efficacy were rated as Low certainty, while knowledge/technical competence, retention, and transfer to practice were rated as Very Low certainty.

Protocol and ethics

No formal review protocol was registered before conducting this systematic literature review. Nevertheless, the review followed a predefined protocol specifying the research questions, eligibility criteria, search strategy, screening procedure, and data extraction process before study selection commenced.

As this study synthesized findings from previously published literature and did not involve the collection of primary data from human participants, ethics committee approval was not required.

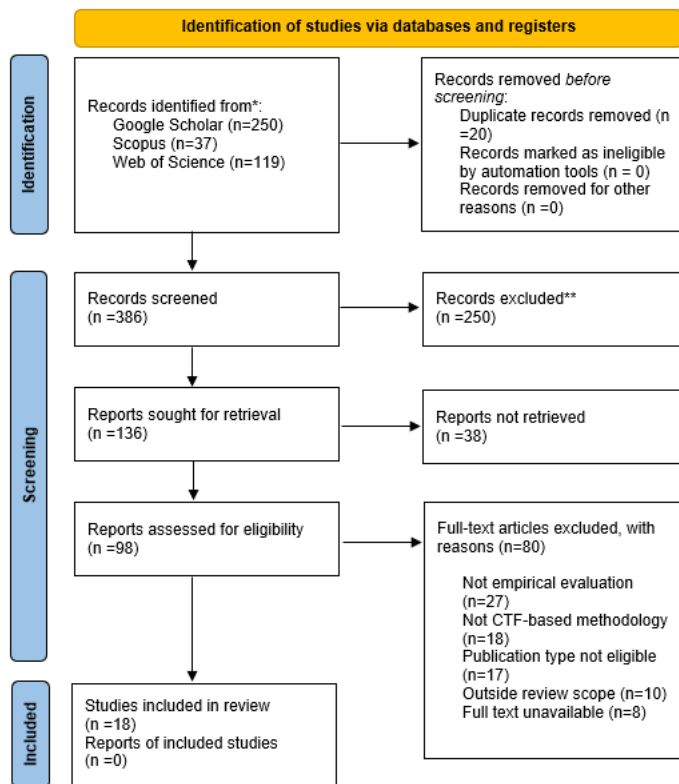
RESULTS

Study Selection

Comprehensive searches across Google Scholar, Scopus, and Web of Science identified 406 records. After duplicate removal and application of the predefined eligibility criteria, 18 empirical studies were included in the qualitative synthesis. The complete database-specific search strategies are presented in Supplementary Table S1 available online at: <https://osf.io/ax24y/files/ahsb4>, while the study selection process is summarised in the PRISMA flow diagram see [Figure 1](#).

Figure 1

PRISMA diagram— study screening and selection

**RQ1: Design and implementation of CTF methodologies*****CTF formats and scenario features (RQ1)***

This section addresses RQ1 by examining how Capture-the-Flag (CTF) methodologies are designed and implemented across different educational contexts. Table 2 summarizes the key characteristics of the 18 studies included in this review. The included studies were published between 2017 and 2025, with a noticeable increase in publication frequency after 2020, indicating growing interest in CTF-based cybersecurity education. Sample sizes varied considerably, ranging from small exploratory pilot studies with as few as six participants (e.g., AR-based CTF implementations) to large-scale outreach programs involving more than 700 participants.

In terms of target populations, higher education students represented the dominant group (11/18 studies, 61%), followed by mixed cohorts that included professionals and secondary-school learners (7/18, 39%). CTF methodologies were implemented both in formal academic settings and in broader professional training and outreach contexts.

The application domains of CTF implementations were diverse. The most commonly represented areas included secure coding and software security (5/18), industrial control systems (ICS) and operational technology (OT) (4/18), and web and IoT security (4/18). A smaller number of studies focused on awareness-oriented topics such as privacy and social engineering (3/18), as well as outreach and pre-college initiatives (2/18). The application domains included CTF methodologies in addressing both technical and non-technical aspects of cybersecurity education see [Table 2](#).

Table 2*Summarizes the key characteristics of the 18 studies included in this review*

Reference	Year	CTF Type	Domain	Context	N
Ford et al.	2017	Offline CTF	General cybersecurity	K-12	36
Antonioli et al.	2017	Attack–Defense	ICS/OT	Academia and industry	Not reported
Schreuders et al.	2018	Jeopardy	Incident response	Higher education	77
Demetrio et al.	2019	Jeopardy	Ethical hacking	Higher education	80–126
Costa et al.	2020	Escape-room	Awareness	Outreach	126

Reference	Year	CTF Type	Domain	Context	N
Gasiba et al.	2020	Design study	Secure coding	Industry	20
Rege & Bleiman	2021	Social engineering	Human factor	Higher education	~25
Sahin et al.	2022	Mixed	Web security	Mixed	103
Ksiezopolski et al.	2022	Jeopardy	Web security	Higher education	~250
Oliveira Junior et al.	2022	Jeopardy	IoT security	Mixed	27
Karampidis et al.	2023	Hybrid	ICS/Industry 4.0	Mixed	41
Collins & Ford	2023	Jeopardy	Secure coding	Higher education	28
Korkiakoski et al.	2023	AR-CTF	Fundamentals	Pilot	6
Larroca et al.	2024	Hybrid	Telecom	Mixed	>100
Ford et al.	2024	Observational	Cybersecurity	Research	11
Vigl & Abramova	2024	Jeopardy	Privacy	Higher education	27
Khoo et al.	2025	Jeopardy	Cybersecurity	Higher education	71
Costa et al.	2025	Hybrid	Gender/STEM	Secondary education	779

Note: Studies are ordered chronologically to illustrate the evolution of CTF-based approaches in cybersecurity education.

The current body of literature on CTF-based learning is summarized by comparing purposes, methods, reported impacts, and limitations across diverse implementations, highlighting both the strengths of CTFs as experiential learning tools and areas requiring further empirical investigation. A detailed comparative analysis of individual studies, including methods, reported effects, and limitations, is provided in Supplementary Table S2, available online at: <https://osf.io/ax24y/files/ahsb4>.

RQ2: Learning outcomes and measurement

This section addresses RQ2 by synthesizing the learning outcomes associated with CTF-based cybersecurity education and examining how these outcomes are operationalized and measured across studies.

Knowledge and technical skills

Improvements in domain-specific knowledge and technical skills were reported in the majority of studies, particularly in higher education and structured training contexts (V. S. Ford, 2017; Ksiezopolski et al., 2022; Vigl & Abramova, 2024). These outcomes were typically reflected in task completion rates, problem-solving performance, and self-reported knowledge gains (Collins & Ford, 2023; Junior et al., 2022). However, in many cases, these findings were based on descriptive statistics or self-report measures rather than validated assessment instruments, limiting the strength of the evidence (Schreuders et al., 2018; Vigl & Abramova, 2024).

Affective outcomes

Affective outcomes, including self-efficacy, confidence, and motivation, were among the most (Schreuders et al., 2018; Vigl & Abramova, 2024). Several studies reported statistically significant improvements in confidence following CTF participation, while others observed positive trends based on Likert-scale responses (Collins & Ford, 2023; V. S. Ford, 2017). High levels of motivation were also reported, particularly in environments incorporating gamification elements and automated feedback systems (Costa et al., 2020; Schreuders et al., 2018). Nevertheless, these outcomes were often measured using non-standardized instruments, which limits comparability across studies (Demetrio et al., 2019).

Engagement and participation

High levels of engagement were reported across most included studies, often reflected in participation rates, submission volumes, and time spent on tasks (Costa et al., 2025; Larroca et al., 2024). Large-scale programs, such as outreach initiatives and competitions, demonstrated particularly strong engagement indicators, including high numbers of submissions and consistent participation across teams (Costa et al., 2025; V. S. Ford, 2017). However, these metrics primarily represent behavioral engagement and do not necessarily correspond to validated measures of learning outcomes (Antonoli et al., 2017; Larroca et al., 2024).

Usability, situational awareness and contextual understanding

Usability and user experience were evaluated in a smaller subset of studies, generally showing positive results (Schreuders et al., 2018). Systems incorporating automation and feedback mechanisms were associated with reduced instructor workload and improved learner experience (Gasiba et al., 2020; Schreuders et al., 2018). Despite

these findings, usability outcomes were rarely linked to measurable learning gains, and their impact on educational effectiveness remains unclear (Gasiba et al., 2020).

Situational awareness and contextual understanding were explored in a limited number of studies, primarily in AR and ICS-based environments (Antonioli et al., 2017; Korkiakoski et al., 2023). These exploratory studies suggest that immersive and system-level scenarios may support the development of real-world recognition skills and contextual reasoning (Korkiakoski et al., 2023). However, the evidence is based on small samples and exploratory designs, limiting the generalizability of these findings (Karampidis et al., 2023; Korkiakoski et al., 2023).

Measurement approaches and outcome synthesis

Overall, the reviewed studies frequently reported positive associations between CTF participation, learner engagement, and affective outcomes, whereas evidence supporting knowledge acquisition and demonstrated competence remained more heterogeneous. Importantly, the lack of standardized measurement instruments and limited use of longitudinal designs restrict the ability to assess retention and transfer of learning outcomes across studies (Ahmed et al., 2024; V. S. Ford, 2017).

RQ3: Methodological quality and evidence strength

This section addresses RQ3 by evaluating the methodological quality, risk of bias, and certainty of evidence across the included studies. Overall, the methodological quality of the reviewed literature was limited, with most studies relying on non-randomized, single-group, descriptive, or exploratory designs.

Study designs and risk of bias

None of the included studies employed randomized controlled trial designs. The majority of studies (14/18, 78%) used single-group pre–post, descriptive, or observational designs, which restrict causal inference and increase susceptibility to bias. In particular, at least 10 of the 18 studies (56%) relied primarily on descriptive or qualitative evaluation approaches without inferential statistical testing, while only a small subset (3/18, 17%) reported any form of statistical testing or effect sizes.

The only nonequivalent-groups experiment (Khoo et al., 2025) illustrates typical threats to internal validity: baseline differences and non-random allocation raised concerns about confounding and selection bias. Although outcome measurement used objective post-test scores with low attrition, the results differed across the two experiments, with null effects before redesign ($p = 0.403$) and a clear post-test advantage after redesign ($p < 0.001$), highlighting sensitivity to task–outcome alignment and assessment quality (Khoo et al., 2025).

Most single-group studies, including Ford et al. (2017), Vigl and Abramova (2024), Karampidis et al. (2023), Demetrio et al. (2019), and Collins and Ford (2023), lacked comparator groups and relied heavily on immediate post-intervention measures. For example, Ford et al. (2017) reported improvements in self-assessed skills (e.g., cryptography knowledge increased from 2.3 to 3.5 on a Likert scale), while Vigl and Abramova (2024) reported a significant increase in confidence ($r = 0.68$, $p < .01$), but both studies lacked long-term follow-up or controlled comparisons. Through these studies, self-reported knowledge, confidence, and motivation were frequently used as outcome indicators, increasing the risk of measurement bias and limiting conclusions about actual competence development.

Platform-centric and field evaluations, such as CyberTrials (Costa et al., 2025), SWaT/S3 (Antonioli et al., 2017), web-application security courses (Ksiezopolski et al., 2022), communications-themed CTFs (Larroca et al., 2024), and IoT-focused CTFs (Junior et al., 2022), primarily relied on participation and performance proxies. For instance, CyberTrials reported 5,220 submitted flags and an average of 43 submissions per team, while Antonioli et al. reported a strong correlation between time and performance ($r \approx 0.97$), but without inferential testing of learning outcomes (Antonioli et al., 2017; Costa et al., 2025; Ksiezopolski et al., 2022).

Qualitative and case-based studies, including hacker-error analyses (I. Ford et al., 2024), social-engineering CTFs (Rege & Bleiman, 2021) and industrial secure-coding challenge design studies (Gasiba et al., 2020), provided useful design insights but limited evidence of learning effectiveness. These studies generally reported sampling strategies and coding procedures, but provided limited detail on reflexivity, triangulation, and analytic transparency. As a result, they are valuable for informing CTF design but cannot support strong causal claims (I. Ford et al., 2024; Gasiba et al., 2020; Rege & Bleiman, 2021).

Study-level quality assessment

To complement the GRADE assessment, a study-level quality appraisal was conducted for each included study using a review-specific rubric. Four methodological domains were evaluated: participant selection, outcome measurement, reporting transparency, and protection against bias. Each domain was rated as High, Moderate, or Low according to the predefined criteria described in Section 3.9. Based on the pattern and importance of these

domain-level ratings, a separate overall methodological quality judgement was subsequently assigned to each study. Ratings were based on study design, participant selection, appropriateness of outcome measures, completeness of methodological reporting, and potential sources of bias. The High, Moderate, and Low ratings refer to the quality of individual methodological domains rather than the overall strength of evidence. A High rating indicates that predefined criteria were met for a specific domain (e.g., reporting transparency or outcome measurement), but does not imply that the study as a whole was at low risk of bias or provided high-certainty evidence. Because most included studies used observational, quasi-experimental, or educational intervention designs, the overall certainty of evidence was assessed separately using the GRADE framework and was generally rated as Low or Very Low. The first reviewer performed the initial appraisal, after which the second reviewer verified all ratings with access to the initial assessments. Any discrepancies were discussed and resolved through consensus. Inter-reviewer agreement statistics were not calculated see [Table 3](#).

Table 3*Study-level quality assessment of included studies*

Study	Study design	Selection	Measurement	Reporting	Protection against bias	Overall quality
Ford et al. (2017)	Single-group pre-post educational intervention	Moderate	Low	Moderate	Low	Low
Antonioli et al. (2017)	Mixed-method / CTF evaluation	Moderate	Low	High	Low	Low
Schreuders et al. (2018)	Mixed methods	Moderate	High	High	Moderate	Moderate
Demetrio et al. (2019)	Educational intervention	Moderate	Low	Moderate	Low	Low
Costa et al. (2020)	Educational intervention	Moderate	Low	Moderate	Low	Low
Gasiba et al. (2020)	Design + expert evaluation	Moderate	Moderate	High	Moderate	Moderate
Rege & Bleiman (2021)	Mixed-method / experience evaluation	Low	Low	High	Low	Low
Sahin et al. (2022)	Mixed methods	Moderate	High	High	Moderate	Moderate
Ksiezopolski et al. (2022)	Observational educational evaluation	Moderate	Moderate	High	Low	Low
Oliveira Junior et al. (2022)	Educational intervention	Moderate	Moderate	Moderate	Low	Low
Karampidis et al. (2023)	Educational intervention	Low	Low	Moderate	Low	Low
Collins & Ford (2023)	Educational intervention	Moderate	Low	High	Low	Low
Korkiakoski et al. (2023)	Pilot study	Low	Low	Moderate	Low	Low
Larroca et al. (2024)	Experience report	Moderate	Low	Moderate	Low	Low
Ford et al. (2024)	Observational study	Moderate	High	High	Moderate	Moderate
Vigl & Abramova (2024)	Educational intervention	Moderate	Moderate	High	Low	Low
Khoo et al. (2025)	Quasi-experimental	Moderate	High	High	Moderate	Moderate
Costa et al. (2025)	Mixed methods	High	Moderate	High	Moderate	Moderate

Outcome-level GRADE assessment

Following GRADE recommendations, certainty was assessed at the outcome level rather than for individual studies. Studies were grouped according to the primary educational outcomes they evaluated, including engagement, confidence/self-efficacy, technical competence, learning retention, and transfer to practice. The certainty ratings therefore represent the overall strength of evidence for each outcome domain rather than the methodological quality of individual studies.

The outcome-level GRADE assessment indicated limited certainty across all five outcome domains. Evidence concerning engagement and motivation and confidence/self-efficacy was rated as Low certainty. Although positive outcomes were frequently reported, the evidence was predominantly derived from uncontrolled educational studies, heterogeneous measurement instruments, and self-reported or indirect indicators.

Evidence concerning knowledge and technical competence was rated as Very Low certainty because of heterogeneous outcome measures, limited use of controlled designs, inconsistent assessment approaches, and frequent reliance on indirect indicators of competence. Evidence for learning retention and transfer to practice was also rated as Very Low certainty, reflecting the scarcity of delayed assessments, longitudinal follow-up, and evidence of application beyond the original training context.

These ratings indicate that frequently reported positive outcomes should not be interpreted as high-certainty evidence of educational effectiveness [Table 4](#)

Table 4

Outcome-level GRADE assessment of the certainty of evidence across included studies

Outcome	Studies (n)	Included studies	Certainty	Reasons for downgrading
Engagement and motivation	12	Demetrio et al. (2019), Schreuders et al. (2018), Ford et al. (2017), Collins & Ford (2023), Costa et al. (2025), Rege & Bleiman (2021), Larroca et al. (2024), Oliveira Junior et al. (2022), Costa et al. (2020), Karampidis et al. (2023), Korkiakoski et al. (2023), Vigl & Abramova (2024)	Low	Predominantly uncontrolled educational studies, self-reported engagement measures, heterogeneous learning environments, risk of publication bias.
Confidence / self-efficacy	7	Ford et al. (2017), Collins & Ford (2023), Costa et al. (2025), Vigl & Abramova (2024), Oliveira Junior et al. (2022), Rege & Bleiman (2021), Korkiakoski et al. (2023)	Low	Mainly questionnaire-based outcomes, limited sample sizes, heterogeneous instruments, few controlled comparisons.
Knowledge and technical competence	10	Ksiezopolski et al. (2022), Gasiba et al. (2020), Sahin et al. (2022), Antonioli et al. (2017), Demetrio et al. (2019), Schreuders et al. (2018), Karampidis et al. (2023), Larroca et al. (2024), Khoo et al. (2025), Oliveira Junior et al. (2022)	Very Low	Diverse outcome measures, lack of randomized or controlled designs, inconsistent assessment methods, indirect measures of competence.
Learning retention	2	Vigl & Abramova (2024), Korkiakoski et al. (2023)	Very Low	Very few studies assessed delayed retention, short follow-up periods, insufficient longitudinal evidence.
Transfer to practice / real-world application	5	Gasiba et al. (2020), Karampidis et al. (2023), Antonioli et al. (2017), Collins & Ford (2023), Schreuders et al. (2018)	Very Low	Evidence largely based on educational settings and qualitative observations, limited workplace validation, indirect measures of transfer.

Note. Certainty ratings were assigned at the outcome level following the GRADE framework. Studies were grouped according to the primary educational outcome they evaluated. Certainty ratings reflect the overall strength of evidence for each outcome rather than the methodological quality of individual studies.

Summary of evidence limitations

Across the reviewed studies, several recurring limitations affected the overall strength of evidence. First, substantial heterogeneity in study designs, contexts, and outcome measures limited comparability and prevented meta-analysis (Ahmed et al., 2024; Demetrio et al., 2019).

Second, most studies relied on single-group pre–post or descriptive designs, which restrict causal inference and increase susceptibility to bias (V. S. Ford, 2017; Karampidis et al., 2023; Vigl & Abramova, 2024).

Outcome measurement was frequently based on self-report instruments or platform-based performance indicators rather than validated assessments of knowledge or skills (Collins & Ford, 2023; Demetrio et al., 2019).

Third, reporting quality was inconsistent. Many studies did not provide effect sizes, variance estimates, or delayed follow-up assessments, limiting the ability to evaluate retention and transfer of learning outcomes (Collins & Ford, 2023; Costa et al., 2020; Schreuders et al., 2018).

Fourth, sample sizes were often small, particularly in exploratory or technology-enhanced CTF implementations, and many studies relied on single-site volunteer cohorts, reducing generalizability (Korkiakoski et al., 2023; Rege & Bleiman, 2021; Sahin et al., 2022).

Finally, performance-based metrics such as leaderboard rankings, submission counts, and time-on-task were widely used but represent indirect proxies of learning. These measures provide insight into engagement but are insufficient to demonstrate competence development without validated outcome measures (Antonoli et al., 2017; Costa et al., 2025; Ksiezopolski et al., 2022).

Taken together, these limitations indicate that the current evidence base is insufficient to support strong conclusions regarding the long-term educational effectiveness of CTF-based learning (Ahmed et al., 2024).

DISCUSSION

This study set out to systematically examine Capture-the-Flag (CTF) methodologies in cybersecurity education by addressing three research questions related to (RQ1) design and implementation, (RQ2) learning outcomes and measurement, and (RQ3) methodological quality and strength of evidence. The findings provide a structured synthesis of a fragmented literature base and highlight recurring patterns in CTF implementation alongside limitations of the current evidence.

The results show that CTF methodologies are implemented across a wide range of formats, contexts, and technological environments. Jeopardy-style CTFs dominate the landscape, accounting for the majority of implementations, particularly in higher education settings.

Jeopardy-style formats were the most common among the included studies, particularly in higher education, whereas Attack–Defense and hybrid formats were less frequently represented and appeared more often in ICS/OT contexts. However, few studies directly compared CTF formats, and the available evidence does not establish whether Jeopardy-style, Attack–Defense, or hybrid formats produce different educational outcomes. Several studies incorporated scaffolding, progressive hints, teamwork, and role specialisation, but these components were rarely isolated experimentally. Their independent effects on learning therefore remain uncertain. VR/AR-based implementations were also identified, although the available evidence was limited and primarily exploratory.

The synthesis of learning outcomes reveals a consistent pattern across studies. CTF participation was frequently associated with positive learner experiences, particularly engagement, confidence, and motivation. However, these findings were predominantly derived from uncontrolled educational studies relying on self-reported outcome measures. Consequently, the magnitude of these effects and their relationship to demonstrated competence remain uncertain. These findings were the most consistently reported across the included studies, although the certainty of evidence remained low according to the GRADE assessment. Evidence for improvements in knowledge and technical skills is more variable. While many studies report gains in problem-solving ability and domain-specific knowledge, these findings are often based on self-reported data or indirect performance indicators such as solve rates and task completion. As a result, it remains difficult to determine the extent to which these improvements reflect actual competence development rather than perceived learning or task-specific familiarity.

A critical issue identified in this review is the lack of consistency in measurement approaches. Most studies rely on non-validated instruments, heterogeneous metrics, and short-term post-intervention assessments. Platform-based indicators, such as leaderboard rankings and submission counts, are frequently used as proxies for learning but do not directly measure knowledge or skill acquisition. Furthermore, very few studies include delayed assessments, limiting insight into retention and transfer of learning outcomes.

Taken together, these findings indicate that while CTF-based learning is commonly reported to support learner engagement and positive learning experiences, the current evidence base is insufficient to support strong claims about long-term knowledge acquisition or transferable skills. The lack of standardized and validated measurement frameworks remains a major barrier to cumulative knowledge building in this field.

The evaluation of methodological quality reveals significant limitations across the reviewed studies. Most research relies on non-randomized, single-group designs, limiting causal inference and increasing susceptibility to bias. In addition, the reliance on self-reported measures and indirect performance indicators further constrains the validity of conclusions.

The GRADE assessment confirmed the limited certainty of the current evidence base, with evidence rated as either Low or Very Low certainty. Evidence for engagement/motivation and confidence/self-efficacy was rated as Low certainty, while evidence for knowledge and technical competence, retention, and transfer to practice was rated as Very Low certainty. Thus, frequently reported positive outcomes should be interpreted cautiously, particularly where they rely on self-reported perceptions or indirect performance indicators.

An important finding of this review is that uncertainty in the current evidence base primarily reflects methodological limitations rather than consistently negative educational outcomes. Although most studies reported favourable learner experiences, comparatively few employed controlled designs, validated assessment instruments, or longitudinal follow-up. Consequently, the available evidence is currently stronger for learner engagement and perceived confidence than for demonstrated competence, long-term retention, or transfer of learning.

However, because these components were rarely isolated experimentally, their independent contribution to learning outcomes cannot be determined from the available evidence. This highlights the need for more rigorous and transparent research designs in future studies.

Despite the methodological limitations identified, the included studies illustrate the use of CTF methodologies as experiential learning activities across diverse cybersecurity education contexts. Several recurring design features—including scaffolding, progressive hints, teamwork, role specialisation, and task–assessment alignment—were observed across the included studies. These features may be considered through established learning theories as possible interpretive frameworks; however, because they were rarely isolated or directly compared, their independent contribution to learning cannot be determined from the available evidence.

From a research perspective, there is a clear need for methodological improvement. Future studies should prioritize comparative designs, including randomized or well-matched control groups, and adopt standardized, validated instruments for measuring knowledge, skills, and self-efficacy. In addition, longitudinal studies are needed to assess retention and transfer of learning outcomes beyond the immediate intervention.

Finally, emerging technologies such as VR/AR and OT-based environments represent developing areas of CTF implementation, but the available evidence remains limited and requires more rigorous evaluation. Without improved methodological consistency and reporting, it will remain difficult to determine under what conditions CTF methodologies lead to sustainable competence development.

LIMITATIONS AND FUTURE RESEARCH

This review has several methodological limitations that should be considered when interpreting the findings. First, the search was restricted to English-language publications identified through Google Scholar, Scopus, and Web of Science. Although these sources provided broad multidisciplinary coverage, relevant studies published in other languages or indexed exclusively in other databases may have been missed. In addition, grey literature, including theses, dissertations, technical reports, preprints, and other non-peer-reviewed sources, was excluded in accordance with the predefined eligibility criteria. While this restriction maintained the review's focus on peer-reviewed empirical research, it may have increased sensitivity to publication bias by reducing the likelihood of identifying unpublished, null, or negative findings.

Second, the review protocol was not prospectively registered. Nevertheless, the review followed predefined research questions, eligibility criteria, search procedures, screening procedures, and data-extraction categories established before study selection. The review was conducted and reported with reference to the PRISMA 2020 reporting framework (Page et al., 2021). Prospective registration would, however, have provided an additional mechanism for documenting methodological decisions and any deviations from the planned review process.

Third, full independent duplicate screening was not conducted for the entire set of retrieved records. The first reviewer screened all records remaining after duplicate removal, while the second reviewer verified records classified as potentially eligible or uncertain. The first reviewer subsequently conducted the full-text eligibility assessment, and the second reviewer verified the full-text eligibility decisions. Because the exact number of records verified by the second reviewer and the number of disagreements were not recorded systematically, inter-reviewer agreement could not be quantified. This procedure may have increased the possibility of erroneous exclusion during the initial screening stage compared with full independent duplicate screening.

Fourth, the use of a review-specific quality appraisal framework represents an additional methodological limitation. The framework enabled a structured assessment across heterogeneous study designs but was not a validated risk-of-bias instrument. Consequently, the domain-level and overall methodological quality ratings involved reviewer judgement despite the use of predefined criteria and verification by a second reviewer. The

resulting study-level quality classifications should therefore be interpreted as structured comparative assessments rather than definitive estimates of methodological quality.

A related limitation concerns the application of the GRADE framework to a heterogeneous educational evidence base. The included studies varied substantially in study design, educational context, outcome definitions, and measurement instruments. Some outcome domains incorporated different forms of evidence, including self-reported perceptions, behavioural indicators, and performance-based measures. Although outcome domains were defined before the certainty assessment, assigning heterogeneous studies to broad categories inevitably involved some reviewer judgement. Misclassification of individual studies or outcomes therefore cannot be completely excluded. In addition, publication bias could not be formally evaluated because of the small number and methodological diversity of studies contributing to individual outcome domains. Consequently, the GRADE ratings should be interpreted as structured judgements about the certainty of evidence within broad educational outcome domains rather than as precise estimates of intervention effects.

In addition to these review-level limitations, limitations within the primary studies constrained the strength of the available evidence. Most studies used non-randomized, single-group pre–post, observational, descriptive, or exploratory designs, limiting causal inference and increasing susceptibility to confounding, selection bias, and measurement bias (Collins & Ford, 2023; Demetrio et al., 2019; V. S. Ford, 2017; Karampidis et al., 2023; Vigl & Abramova, 2024). Outcome measurement frequently relied on self-reported knowledge, confidence, motivation, or engagement, while validated assessment instruments were uncommon (Collins & Ford, 2023; Demetrio et al., 2019; V. S. Ford, 2017; Vigl & Abramova, 2024). These methodological characteristics contributed to the Low or Very Low certainty ratings assigned to the outcome domains.

Performance-based measures also require cautious interpretation. Several studies used indicators such as leaderboard position, flag submissions, task-completion rates, solve rates, or time-based performance measures (Antonioli et al., 2017; Costa et al., 2025; Ksiezopolski et al., 2022; Larroca et al., 2024). Although these measures provide useful information about participation and task performance, they should be regarded as indirect indicators of learning unless they are linked to explicit and validated assessments of cybersecurity knowledge or competence.

Reporting quality and sample characteristics further constrained the evidence base. Effect sizes, variance estimates, and delayed follow-up assessments were inconsistently reported across the included studies (Collins & Ford, 2023; Costa et al., 2020; Schreuders et al., 2018). Several studies involved relatively small, single-site, or volunteer samples, particularly exploratory and technology-enhanced implementations (Collins & Ford, 2023; Korkiakoski et al., 2023; Rege & Bleiman, 2021; Sahin et al., 2022). These characteristics reduced comparability and generalisability and contributed to uncertainty regarding whether reported short-term outcomes persisted beyond the immediate training context.

The substantial heterogeneity of CTF implementations also limited synthesis. The included studies covered Jeopardy-style, Attack–Defense, hybrid, immersive, and other CTF formats across K–12, higher education, professional training, outreach, and ICS/OT contexts (Antonioli et al., 2017; Costa et al., 2025; Demetrio et al., 2019; V. S. Ford, 2017; Gasiba et al., 2020; Karampidis et al., 2023; Korkiakoski et al., 2023). Few studies directly compared different CTF formats or experimentally isolated individual design components. Consequently, the present review cannot determine whether particular CTF formats or design features, such as scaffolding, progressive hints, teamwork, or role specialisation, produce different educational outcomes.

Evidence concerning learning retention and transfer was particularly limited. Most studies assessed outcomes during or immediately after CTF participation, while delayed assessments were uncommon. This restricts conclusions about whether knowledge, technical competence, confidence, or other learning outcomes are sustained over time or transferred to new tasks and professional contexts. Accordingly, the certainty of evidence for retention and transfer to practice was rated as Very Low.

Future research should address these limitations through more rigorous and comparable study designs. Randomized studies may be appropriate in some educational contexts, while carefully matched comparison groups and stronger quasi-experimental designs may provide more feasible alternatives in authentic educational and professional training environments. Greater use of shared and validated measures of cybersecurity knowledge, technical competence, confidence, and self-efficacy would improve cross-study comparability. Objective performance measures should complement self-reported outcomes rather than be treated as interchangeable with them.

Longitudinal assessment is particularly important. Future studies should include delayed post-tests to determine whether learning outcomes are retained beyond the immediate intervention and whether knowledge and skills acquired in CTF environments transfer to new tasks, workplace activities, or other authentic cybersecurity contexts. Multi-site studies with adequately powered samples would further strengthen generalisability and enable investigation of differences between learner populations and educational settings.

Future evaluations of OT/ICS and immersive CTF environments should combine realistic scenarios with clearly defined learning outcomes and appropriate objective performance measures. Component-level studies are

also needed to determine whether specific design features—such as scaffolding, progressive hints, teamwork, role specialisation, or immersive technologies—contribute independently to learning. More consistent reporting of effect sizes, variance estimates, intervention characteristics, and follow-up outcomes would facilitate future quantitative synthesis and provide a stronger evidence base for determining whether, when, and under what conditions CTF-based learning contributes to sustained cybersecurity competence.

COMPARISON WITH PRIOR WORK

Prior research on CTF-based learning has primarily focused on Jeopardy-style formats in undergraduate education, with an emphasis on descriptive engagement outcomes rather than standardized measures of learning (Švábenský et al., 2021; Vykopal et al., 2020). While concepts such as self-efficacy have been implicitly addressed through mechanisms like competition and feedback, they have rarely been measured using validated instruments. Similarly, theoretical constructs such as scaffolding and cognitive load are reflected in CTF design practices but are seldom explicitly operationalized or empirically tested.

This lack of theoretical integration and measurement consistency represents a key gap in literature. Prior studies have rarely examined how specific design features—such as hint systems, task progression, or team structures—relate to learning outcomes, nor have they systematically evaluated higher-order competencies such as situational awareness or decision-making.

The present review extends prior work by integrating evidence across 18 studies using the GRADE framework to assess the strength and certainty of findings. The synthesis highlights that while engagement and short-term confidence gains are consistently reported, the overall certainty of evidence remains low due to methodological limitations, including reliance on self-reported measures and single-group designs. In contrast to earlier descriptive reviews, this study provides a structured evaluation of both learning mechanisms and evidence quality, thereby clarifying where CTF methodologies demonstrate promise and where further empirical validation is required.

CONCLUSION

This study presented a systematic review of Capture-the-Flag (CTF) methodologies in cybersecurity education, synthesizing evidence across 18 empirical studies published between 2017 and 2025. By addressing three research questions related to design and implementation, learning outcomes, and methodological quality, the review provides a structured overview of how CTF-based learning is currently applied and evaluated.

The findings show that CTF methodologies have been implemented as experiential learning activities across diverse educational contexts, with Jeopardy-style formats being the most frequently represented in the included studies. More complex formats, such as Attack–Defense and scenario-based CTFs, were less frequently examined, and the available evidence does not establish whether these formats produce different educational outcomes. Across implementations, design features such as scaffolding, teamwork, progressive hints, and progressive difficulty were commonly incorporated into CTF activities. However, these features were rarely isolated or directly compared, and their independent contribution to educational outcomes therefore remains uncertain.

Across the reviewed studies, outcomes related to learner engagement, motivation, and perceived confidence were frequently reported. However, these findings were predominantly derived from uncontrolled studies using heterogeneous outcome measures, self-reported instruments, or indirect indicators of performance. Consequently, the magnitude, consistency, and educational significance of these reported outcomes remain uncertain. The methodological quality assessment further demonstrated important limitations in the current evidence base. Most studies employed non-randomized, single-group, observational, or exploratory designs, and very few included longitudinal assessments or validated measures of demonstrated competence.

The outcome-level GRADE assessment reflected these methodological limitations. Evidence for engagement/motivation and confidence/self-efficacy was rated as Low certainty, whereas evidence for demonstrated knowledge and technical competence, learning retention, and transfer to practice was rated as Very Low certainty. These findings indicate that frequently reported engagement and perceived confidence should not be interpreted as equivalent to demonstrated or sustained cybersecurity competence.

Despite these limitations, this review contributes to the field in three ways. First, it provides a structured mapping of CTF methodologies across formats, contexts, and domains. Second, it synthesizes reported learning outcomes while critically examining how these outcomes have been operationalized and measured. Third, it evaluates both study-level methodological quality and outcome-level certainty of evidence, providing a transparent assessment of the strengths and limitations of the current research base.

Across the included studies, CTF participation was frequently associated with learner engagement and perceived confidence; however, the certainty of evidence for these outcomes was Low. Evidence for demonstrated knowledge and technical competence, learning retention, and transfer to practice was rated Very Low. The available evidence therefore does not support conclusions about which CTF format or individual design feature produces

superior educational outcomes. More rigorous comparative studies, validated outcome measures, and longitudinal evaluations are required before robust conclusions can be drawn regarding the effectiveness of CTF-based learning in developing sustained cybersecurity competence.

Acknowledgement

The authors would like to acknowledge the support of Tallinn University of Technology (TalTech), Østfold University College, and Albstadt-Sigmaringen University. The authors also acknowledge the support provided through the ATHENA projects, which contributed to the broader research context in which this systematic literature review was conducted.

Funding

This study was funded by the Digital Europe Programme: Project ATHENA (grant number 101127970). This study was funded by the Digital Europe Programme: Project CHESS (grant number 101087529).

Ethical statement

This study is a systematic literature review based exclusively on previously published research. No human participants were recruited, no personal data were collected, and no experimental procedures were conducted. Therefore, institutional review board (IRB) or ethics committee approval and informed consent were not required.

Competing interests

The authors declare that they have no competing financial or non-financial interests that could have influenced the work reported in this article.

Author contributions

T.M: Conceptualization, methodology, literature search, data collection, data analysis and synthesis, visualization, and writing – original draft. R.G.L: Supervision, validation, writing – review and editing, and critical feedback. S.S: Supervision, writing – review and editing, and critical feedback. All authors reviewed the manuscript, provided comments and suggestions for improvement, and approved the final version for publication.

Data availability

This systematic literature review is based on previously published studies cited in the article. The data supporting the findings of this review were derived from these publicly available publications. The extracted and synthesized data used in the review are presented within the article and its supplementary materials. Additional materials related to the review process may be made available by the corresponding author upon reasonable request.

AI disclosure

The author acknowledges the use of AI-assisted tools (e.g., ChatGPT) to support translation and language editing of the manuscript. These tools were used solely to improve clarity, grammar, and readability.

Biographical sketch

Triin Muulmann is a PhD student and Junior Researcher at the Estonian Maritime Academy, Tallinn University of Technology (TalTech), Estonia. She holds a master's degree and has extensive experience in vocational IT and cybersecurity education. Her primary research field is cybersecurity education, with a particular focus on innovative and practice-oriented learning methods. Her research interests include Capture the Flag (CTF)-based learning, virtual reality and immersive learning environments, cybersecurity self-efficacy, learning retention, and the development of practical cybersecurity competencies in vocational and higher education.

Ricardo Gregorio Lugo is a Senior Researcher at the Estonian Maritime Academy, Tallinn University of Technology (TalTech), Estonia, and is also affiliated with Østfold University College, Norway. He holds a PhD, with doctoral research focusing on micro- and macrocognitive factors of performance in cyber defence. His primary research expertise is in experimental and applied psychology and human factors. His research interests include human factors in cybersecurity, cognitive performance, cyber threat identification, decision-making, self-efficacy, and the role of education and experience in cybersecurity performance.

Stefan Sütterlin is affiliated with Albstadt-Sigmaringen University, Germany, and Østfold University College, Norway. His research expertise spans psychology, human factors, and cybersecurity, with a particular emphasis on understanding human behaviour and cognitive processes in digital and security-related environments. His research interests include human factors in cybersecurity, cyber threat identification and decision-making, cognitive and behavioural aspects of security, and the interaction between psychological factors, education, and cybersecurity performance.

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of Lectito Publications and/or the editor(s). Lectito Publications and/or the editor(s) disclaim responsibility for any injury to persons or property resulting from any ideas, methods, instructions, or products referred to in the content.

REFERENCES

- Ahmed, A., Watterson, C., Alhashmi, S., & Gaber, T. (2024). How universities teach cybersecurity courses online: A systematic literature review. *Frontiers in Computer Science*, 6, Article 1499490. <https://doi.org/10.3389/fcomp.2024.1499490>
- Alnajim, A. M., Habib, S., Islam, M., AlRawashdeh, H. S., & Wasim, M. (2023). Exploring cybersecurity education and training techniques: A comprehensive review of traditional, virtual reality, and augmented reality approaches. *Symmetry*, 15(12), Article 2175. <https://doi.org/10.3390/sym15122175>
- Antonioli, D., Ghaeini, H. R., Adepu, S., Ochoa, M., & Tippenhauer, N. O. (2017). Gamifying education and research on ICS security: Design, implementation and results of S3. In *Proceedings of the 2017 Workshop on Cyber-Physical Systems Security and Privacy* (pp. 93–102). Association for Computing Machinery. <https://doi.org/10.1145/3140241.3140253>
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.
- Barker, T. H., Habibi, N., Aromataris, E., Stone, J. C., Leonardi-Bee, J., Sears, K., Hasanoff, S., Klugar, M., Tufanaru, C., Moola, S., & Munn, Z. (2024). The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. *JBIE Evidence Synthesis*, 22(3), 378–388. <https://doi.org/10.11124/JBIES-23-00268>
- Collins, J., & Ford, V. (2023). Teaching by practice: Shaping secure coding mentalities through cybersecurity CTFs. *Journal of Cybersecurity Education, Research and Practice*, 2022(2). <https://doi.org/10.32727/8.2023.8>
- Costa, G., De Francisci, S., Renieri, M., & Valiani, S. (2025). Tackling the gender gap in cybersecurity education. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1* (pp. 234–240). Association for Computing Machinery. <https://doi.org/10.1145/3641554.3701807>
- Costa, G., Lualdi, M., Ribaud, M., & Valenza, A. (2020). A nerd dogma: Introducing CTF to non-expert audience. In *Proceedings of the 21st Annual Conference on Information Technology Education* (pp. 413–418). Association for Computing Machinery. <https://doi.org/10.1145/3368308.3415405>
- Demetrio, L., Lagorio, G., Ribaud, M., Russo, E., & Valenza, A. (2019). ZenHackAdemy: Ethical hacking @ DIBRIS. In *Proceedings of the 11th International Conference on Computer Supported Education* (Vol. 1, pp. 405–413). SciTePress.
- European Cyber Security Organisation. (2020). *Understanding cyber ranges: From hype to reality*. https://www.evanssion.com/wp-content/uploads/2025/02/ECSO_Paper_on_cyber_ranges_understanding-cyber-ranges-from-hype-to-reality_2.pdf
- European Union Agency for Cybersecurity. (2018). *Cybersecurity culture guidelines: Behavioural aspects of cybersecurity*.
- Ford, I., Soneji, A., Kokulu, F. B., Vadayath, J., Basque, Z. L., Vipat, G., & Shoshitaishvili, Y. (2024). "Watching over the shoulder of a professional": Why hackers make mistakes and how they fix them. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 350–368). IEEE. <https://doi.org/10.1109/SP54263.2024.00202>
- Ford, V. S. (2017). Capture the flag unplugged: An offline cyber competition. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (pp. 225–230). Association for Computing Machinery. <https://doi.org/10.1145/3017680.3017783>
- Gasiba, T., Lechner, U., Pinto-Albuquerque, M., & Zouitni, A. (2020). Design of secure coding challenges for cybersecurity education in the industry. In *International Conference on the Quality of Information and Communications Technology* (pp. 223–237). Springer. https://doi.org/10.1007/978-3-030-58793-2_18
- Gervais, J. (2016). The operational definition of competency-based education. *The Journal of Competency-Based Education*, 1(2), 98–106. <https://doi.org/10.1002/cbe2.1011>
- GRADE Working Group. (2025). GRADE. <https://www.gradeworkinggroup.org/>

- Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, 34(4), 285–291. <https://doi.org/10.3233/EFI-180221>
- İpek, Z. H., Gözümlü, A. İ. C., Papadakis, S., & Kalogiannakis, M. (2023). Educational applications of the ChatGPT AI system: A systematic review research. *Educational Process: International Journal*, 12(3), 26–55. <https://doi.org/10.22521/edupij.2023.123.2>
- Junior, A. O., Funchal, G., Queiroz, J., Loureiro, J., Pedrosa, T., Parra, J., & Leitão, P. (2022). Learning cybersecurity in IoT-based applications through a capture the flag competition. In *2022 IEEE 20th International Conference on Industrial Informatics (INDIN)* (pp. 560–565). IEEE. <https://doi.org/10.1109/INDIN51773.2022.9976079>
- Kalogiannakis, M., Papadakis, S., & Zourmpakis, A.-I. (2021). Gamification in science education: A systematic review of the literature. *Education Sciences*, 11(1), Article 22. <https://doi.org/10.3390/educsci11010022>
- Karampidis, K., Panagiotakis, S., Vasilakis, M., Lamari, A. T., Markakis, E., & Papadourakis, G. (2023). Digital training for cybersecurity in industrial fields via virtual labs and capture-the-flag challenges. In *2023 32nd Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE)* (pp. 1–6). IEEE. <https://doi.org/10.23919/EAEEIE55804.2023.10181644>
- Khoo, L. J., Yatim, M. H. M., & Wong, Y. S. (2025). Research on capture the flag exercises for cybersecurity skill training among Malaysian undergraduates. *Journal of Human Centered Technology*, 4(1), 1–9. <https://doi.org/10.11113/humentech.v4n1.87>
- Korkiakoski, M., Antila, A., Annamaa, J., Sheikhi, S., Alavesa, P., & Kostakos, P. (2023). Hack the room: Exploring the potential of an augmented reality game for teaching cyber security. In *Proceedings of the Augmented Humans International Conference 2023* (pp. 349–353). Association for Computing Machinery. <https://doi.org/10.1145/3582700.3583955>
- Ksiezopolski, B., Mazur, K., Miskiewicz, M., & Rusinek, D. (2022). Teaching a hands-on CTF-based web application security course. *Electronics*, 11(21), Article 3517. <https://doi.org/10.3390/electronics11213517>
- Larroca, F., Belcredi, G., García, R., González, G. G., Inglés, L., Mariño, C., & Randall, M. (2024). Building a radio community across students, industry and enthusiasts with capture the flag. In *2024 XVI Congreso de Tecnología, Aprendizaje y Enseñanza de la Electrónica (TAAEE)* (pp. 1–6). IEEE. <https://doi.org/10.1109/TAAEE59541.2024.10604950>
- Lazarov, W., Schaefitel-Tähtinen, T., Squillace, J., Martinasek, Z., Coufalikova, A., Helenius, M., & Fujdiak, R. (2025). Lessons learned from using cyber range to teach cybersecurity at different levels of education. *Technology, Knowledge and Learning*. Advance online publication. <https://doi.org/10.1007/s10758-025-09840-y>
- Leune, K., & Petrilli, S. J., Jr. (2017). Using capture-the-flag to enhance the effectiveness of cybersecurity education. In *Proceedings of the 18th Annual Conference on Information Technology Education* (pp. 47–52). Association for Computing Machinery. <https://doi.org/10.1145/3125659.3125686>
- Malone, M., Wang, Y., James, K., Anderegg, M., Werner, J., & Monroe, F. (2021). To gamify or not? On leaderboard effects, student engagement and learning outcomes in a cybersecurity intervention. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 1135–1141). Association for Computing Machinery. <https://doi.org/10.1145/3408877.3432544>
- Mulder, M. (2014). Conceptions of professional competence. In S. Billett, C. Harteis, & H. Gruber (Eds.), *International handbook of research in professional and practice-based learning* (pp. 107–137). Springer.
- National Institute of Standards and Technology. (2025). *Information Technology Laboratory: Computer Security Resource Center*. <https://csrc.nist.gov/>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, Article n71. <https://doi.org/10.1136/bmj.n71>
- Rege, A., & Bleiman, R. (2021). Collegiate social engineering capture the flag competition. In *2021 APWG Symposium on Electronic Crime Research (eCrime)* (pp. 1–11). IEEE.
- Sahin, M., Ünlü, T., Hébert, C., Shepherd, L. A., Coull, N., & Mc Lean, C. (2022). Measuring developers' web security awareness from attack and defense perspectives. In *2022 IEEE Security and Privacy Workshops (SPW)* (pp. 31–43). IEEE. <https://doi.org/10.1109/SPW54247.2022.9833858>
- Schreuders, Z. C., Shaw, T., Mac Muireadhaigh, A., & Staniforth, P. (2018). Hackerbot: Attacker chatbots for randomised and interactive security labs, using SecGen and oVirt. In *2018 USENIX Workshop on Advances in*

- Security Education (ASE 18). USENIX Association.
<https://www.usenix.org/conference/ase18/presentation/schreuders>
- Stouffer, K., Pease, M., Tang, C., Zimmerman, T., Pillitteri, V., Lightman, S., Hahn, A., Saravia, S., Sherule, A., & Thompson, M. (2023). *Guide to operational technology (OT) security* (NIST Special Publication 800-82r3). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-82r3>
- Švábenský, V., Čeleda, P., Vykopal, J., & Brišáková, S. (2021). Cybersecurity knowledge and skills taught in capture the flag challenges. *Computers & Security*, 102, Article 102154.
<https://doi.org/10.1016/j.cose.2020.102154>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Vigl, W., & Abramova, S. (2024). Design and use of privacy capture-the-flag challenges in an introductory class on information privacy and security. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1* (pp. 618–624). Association for Computing Machinery.
<https://doi.org/10.1145/3649217.3653572>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Vykopal, J., Švábenský, V., & Chang, E.-C. (2020). Benefits and pitfalls of using capture the flag games in university courses. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 752–758). Association for Computing Machinery. <https://doi.org/10.1145/3328778.3366893>
- Zouahi, H. (2023). Gamifying cybersecurity education: A CTF-based approach to engaging students in software security laboratories. *Proceedings of the Canadian Engineering Education Association (CEEAA)*.
<https://doi.org/10.24908/pceea.2023.17071>