

ACCELERATING THE LEARNING CURVE: GENERATIVE AI FOR WORKED EXAMPLES IN HUMAN FACTORS ANALYSIS

Ariane Reisier[1,2] & Moacyr Machado Cardoso Júnior[1]

[1]Technological Institute of Aeronautics - ITA
[2]PhD. Candidate

Abstract

This paper presents a GenAI-assisted worked examples approach for teaching complex human factors analysis methodologies in graduate-level aviation safety education. Comprehensive worked examples were developed for seven HFE methods — Hierarchical Task Analysis (HTA), Goal-Directed Task Analysis (GDTA), Operator-Event Sequence Diagram (OESD), Event Analysis of Systemic Teamwork (EAST), Critical Decision Method (CDM), Conceptual Mapping, and Cognitive Work Analysis (CWA) — using authentic aviation operational scenarios. Following implementation in course MP-261 at Instituto Tecnológico de Aeronáutica (ITA), 15 of 21 enrolled graduate students completed a structured satisfaction questionnaire. Inferential analyses confirmed strongly positive reactions: templates were perceived as facilitating HFE inclusion in projects (Q5: mean = 4.60, $p < .001$, $r = 0.882$) and reducing methodological complexity (Q6: mean = 4.53, $p < .001$, $r = 0.874$); 93.3% of respondents reported easier learning relative to other technical courses ($p < .001$); and the recommendation rate was near-universal (Q10: mean = 4.80, $p < .001$, $r = 0.929$). HTA and EAST received the strongest per-methodology ratings, while Conceptual Mapping and GDTA elicited more variable responses, consistent with Cognitive Load Theory predictions. These results provide statistically grounded, Level 1 evidence supporting GenAI-generated worked examples as a promising and scalable pedagogical strategy for safety-critical engineering education.

Keywords: Human Factors Engineering, Generative AI, Aviation Safety Education, Worked Examples, Cognitive Load Theory

1. Introduction)

Aviation safety stands at the intersection of technological sophistication and human performance. Modern aircraft systems, air traffic management infrastructures, and maintenance organizations are among the most complex sociotechnical systems ever created, demanding that practitioners simultaneously command engineering principles, operational knowledge, and a rigorous understanding of human behavior under high workload and uncertainty. This complexity has driven the adoption of Human Factors Engineering (HFE) as a mandatory discipline within regulatory frameworks worldwide: EASA Part-145, FAR Part 5, and ICAO Annex 19 all mandate systematic human factors analysis as a cornerstone of Safety Management Systems (SMS) [1]. In Brazil, this demand is particularly salient. The growth of Embraer's commercial and executive product lines, the expansion of Brazilian airspace operations coordinated by DECEA, and the consolidation of maintenance, repair, and overhaul (MRO) capacity have created acute demand for engineers who can perform rigorous HFE analyses. Graduate programs in aerospace engineering—including the Graduate Program in Space Sciences and Technologies (CTE) at Instituto Tecnológico de Aeronáutica (ITA)—have responded by incorporating HFE methods into their curricula. However, the gap between regulatory expectation and educational delivery remains substantial. The methods demanded by industry and regulators—Hierarchical Task Analysis (HTA), Goal-Directed Task Analysis (GDTA), Operator-Event Sequence Diagrams (OESD),

Cognitive Work Analysis (CWA), and related frameworks—are analytically sophisticated tools that require students to integrate procedural methodology with deep domain knowledge [2]. The pedagogical challenge is not merely one of content transfer, but of enabling students to think systemically about human-machine interaction within safety-critical operational contexts.

Research in educational psychology consistently identifies four barriers that make HFE methods particularly difficult to teach at graduate level: (1) Cognitive Complexity—students must simultaneously manage procedural knowledge, domain expertise, and analytical reasoning [2]; (2) Domain Knowledge Deficit—effective application requires deep understanding of aviation operations that novice students lack; (3) Abstract Methodology—without concrete examples, students struggle to connect theory to practice; and (4) Limited Exposure—traditional instruction provides insufficient exemplars for pattern recognition and skill development. Cognitive Load Theory (CLT) provides the theoretical basis for addressing these barriers [3, 4]. CLT distinguishes between intrinsic cognitive load (the inherent complexity of the material), extraneous load (load imposed by poor instructional design), and germane load (cognitive effort directed toward schema formation). Traditional instruction of HFE methods tends to maximize extraneous load by presenting tools in abstract form without sufficient worked examples, thereby leaving insufficient cognitive resources for the deep schema formation needed for genuine competence. Worked examples—detailed, annotated demonstrations of the complete problem-solving process—are the most empirically validated intervention for reducing extraneous load in complex technical domains [3, 5]. By providing the solution structure, worked examples allow learners to allocate cognitive resources to understanding the underlying rationale rather than constructing solutions from scratch. The evidence base for this approach spans mathematics, physics, medicine, and engineering education [5, 6], but its application to HFE methods in aviation safety education has not been systematically studied. Recent advances in Generative Artificial Intelligence (GenAI) create a novel opportunity to address the scale problem inherent in worked-example-based instruction. Large Language Models (LLMs) such as Claude (Anthropic), GPT-4 (OpenAI), and Gemini (Google) can generate contextually rich, methodologically detailed worked examples across diverse aviation scenarios at a pace and volume impossible for individual instructors to match [7]. Crucially, these systems can adapt scenarios to specific regulatory contexts, operational roles, and learning objectives—properties essential for HFE education, where contextual authenticity directly affects analytical validity. However, the mere availability of AI-generated content does not guarantee pedagogical effectiveness. Student satisfaction—defined in the Kirkpatrick model [8] as Level 1 evaluation—represents a necessary (if not sufficient) condition for learning: students who perceive an instructional approach as unhelpful are unlikely to engage with it deeply. Systematic evidence on student satisfaction with AI-generated worked examples in specialized aviation HFE education is conspicuously absent from the literature. This study addresses that gap.

1.1 Research Objectives

Primary Objective: To assess student satisfaction with GenAI-developed worked examples as a learning strategy for human factors analysis methodologies in graduate-level aviation safety education.

Specific Objectives:

a. Develop comprehensive worked examples for seven human factors analysis methodologies: Hierarchical Task Analysis (HTA), Goal-Directed Task Analysis (GDTA), Operator-Event Sequence Diagram (OESD), Event Analysis of Systemic Teamwork (EAST), Critical Decision Method (CDM), Conceptual Mapping, and Cognitive Work Analysis (CWA, covering all five analytical phases)—using GenAI assistance, covering authentic aviation operational scenarios

b. Implement a worked-example-based pedagogical approach in a graduate-level human factors course (MP-261) at ITA.

c. Measure student satisfaction with the worked examples for each methodology using a post-intervention questionnaire.

d. Identify which methodologies benefited most from the worked-example approach, and characterize perceived strengths and areas for improvement.

e. Provide evidence-based recommendations for the integration of GenAI-generated worked examples into safety-critical engineering education curricula.

2. Theoretical Framework

2.1 Cognitive Load Theory and the Worked Example Effect

Cognitive Load Theory, originally formulated by Sweller [3] and subsequently extended by van Merriënboer and Paas [4], provides the dominant theoretical lens for understanding skill acquisition in complex technical domains. CLT proposes that working memory has a severely limited capacity, and that effective instructional design must minimize unnecessary load on working memory while maximizing the formation of long-term memory schemas. The theory distinguishes three load types: intrinsic load, determined by the inherent complexity of the content and the learner's prior knowledge; extraneous load, determined by the quality of instructional design; and germane load, which reflects the cognitive effort directed toward schema formation and automation. HFE analysis methods inherently carry high intrinsic load because they require simultaneous processing of procedural rules, domain-specific knowledge, and inferential reasoning about human behavior in complex systems [2]. Traditional instructional approaches—presenting the method's formal specification followed by an unguided application exercise—impose high extraneous load because students must simultaneously discover the solution procedure, manage domain knowledge gaps, and apply the analytical framework. This triply loaded condition is precisely the configuration that CLT identifies as most likely to overwhelm working memory and inhibit learning. Worked examples address this configuration directly. By presenting the complete solution path with explanatory annotations, they eliminate the extraneous load of solution-generation, freeing cognitive resources for comprehension of the underlying rationale. The "worked example effect"—the reliable finding that worked examples produce superior learning outcomes compared to equivalent problem-solving practice, particularly for novice learners has been replicated across diverse domains and age groups [5, 6]. It was further demonstrated that the effectiveness of worked examples depends critically on the quality of self-explanation prompts embedded within them: examples that explicitly highlight the rationale for each analytical step produce deeper learning than examples presenting only the steps themselves [9].

2.2 Human Factors Analysis Methods in Engineering Education

The challenge of teaching HFE analysis methods is well-documented in the safety engineering literature. A catalogue of 107 human factors methods are presented by Stanton et al. [2] and note that their effective application requires not only procedural competence but also "practitioner wisdom" - the ability to select the appropriate method for a given problem, interpret results in operational context, and communicate findings to multidisciplinary teams. This kind of wisdom is accumulated through extensive practice with diverse cases, which traditional instruction cannot provide at scale.

Research on HTA pedagogy has highlighted the challenge of teaching hierarchical decomposition to students who lack operational mental models [10]. Annett and Duncan note that the apparent simplicity of HTA's notation belies the analytical sophistication required for meaningful task decomposition, particularly in domains with complex operator-system interactions [10]. The applicability of this challenge to augmented-reality and mixed-initiative health environments was subsequently demonstrated by [11], who showed that HTA-based frameworks remain essential for identifying human error in sequential tasks even as automation levels increase. For Goal-Directed Task Analysis, Endsley [12] and Endsley and Jones [13] note that professionals consistently underestimate the cognitive demands of Situation Awareness, tending to conflate Level 1 (Perception) with the higher-order Level 2 (Comprehension) and Level 3 (Projection) processes that constitute expert SA. This conflation leads to analytically shallow GDTA outputs that enumerate perceptual cues without addressing the integration and predictive reasoning that define operational expertise. Worked examples that explicitly annotate SA-level distinctions are therefore particularly valuable for GDTA instruction. Cognitive Work Analysis has received specific attention as a particularly demanding framework to teach. Vicente [14] acknowledges that CWA's breadth—spanning five analytical phases from Work Domain Analysis through Worker Competencies Analysis—poses substantial pedagogical challenges. [15] note that students frequently struggle to maintain analytical coherence across CWA's phases, losing

the systemic thread that connects the abstraction hierarchy to operator strategies. Template-based approaches that structure each analytical phase have been proposed as a remedy, but their effectiveness in aviation engineering education had not been empirically evaluated prior to the present study. Shorrock and Williams [16] advocate for case-based learning approaches that ground method application in realistic accident and incident scenarios, arguing that contextual authenticity is a necessary condition for meaningful HFE analysis. This principle is reinforced by [17], who emphasize that effective HFE education must develop students' mental models of operational realities, not merely their knowledge of analytical procedures. Both recommendations converge on the worked-example approach: high-quality worked examples based on authentic aviation scenarios simultaneously provide methodological guidance and operational context, addressing both the procedural and the domain-knowledge deficits that impede novice analysts.

2.3 Generative AI in Engineering Education

The application of Large Language Models in higher education is a rapidly evolving research area. Mollick and Mollick [7] provide one of the most systematic frameworks for this application, identifying seven distinct pedagogical roles for AI systems: tutor, coach, mentor, teammate, tool, simulator, and student. Their analysis suggests that the "AI as simulator" and "AI as tool" roles are particularly well-suited to technical domains, where the ability to generate diverse, contextually authentic practice scenarios can substantially accelerate skill acquisition. This framework directly informs the present study's use of GenAI for worked-example generation: the AI functions simultaneously as a tool (automating scenario authoring) and as a simulator (providing realistic operational contexts for HFE analysis). Kasneci et al. [18] provide a comprehensive review of LLMs in education, noting that these systems are particularly effective for generating worked examples and explanatory content in structured domains. They highlight the critical importance of pedagogical scaffolding in AI-generated content: raw LLM outputs, while often accurate, typically lack the motivational framing, self-explanation prompts, and error analysis that characterize effective instructional design. Their review underscores the need for educator-guided prompt engineering to produce educationally effective AI-generated content—a need addressed in the present study through a structured development protocol (Section 3.3). Dai et al. [19] found that undergraduate engineering students rated AI-generated worked examples as significantly more helpful than traditional textbook examples for complex multi-step problems, attributing this preference to the AI examples' more explicit step-by-step reasoning and contextual annotations. However, they also noted that student satisfaction varied substantially with perceived authenticity: worked examples grounded in realistic scenarios were rated significantly more helpful than those using artificial or oversimplified contexts. This finding has direct implications for the present study's design, motivating the selection of authentic aviation operational scenarios for all worked examples. Within the aviation safety domain, [11] demonstrated the value of structured analytical frameworks for human error identification in sequential aviation tasks, establishing a methodological precedent for the integration of AI-assisted tools in HFE analysis. The present study extends this line of work by examining whether GenAI-generated worked examples—rather than AI tools applied directly to analysis—can accelerate the development of HFE analytical competence in graduate students.

2.4 Student Satisfaction as an Evaluation Criterion

The Kirkpatrick model of training evaluation [8] defines four levels of assessment: Reaction (Level 1), Learning (Level 2), Behavior (Level 3), and Results (Level 4). While higher-level evaluations are ultimately more consequential, Level 1 satisfaction data serve important functions in formative evaluation: they identify perceived barriers to engagement, surface design issues that may suppress learning, and provide proxies for learner motivation—a variable that substantially moderates the worked example effect [9]. Alliger et al. [20] meta-analysis confirmed that utility reactions—students' perceptions of how useful a training intervention is for their work—are significantly correlated with transfer of learning, validating satisfaction measurement as a meaningful pedagogical indicator in technical domains. In the context of this study, satisfaction is operationalized along two dimensions: (a) perceived facilitation—the extent to which worked-example templates reduced the mechanical complexity of method application and allowed focus on analytical essentials; and (b) comparative value—students'

judgment of the worked-example approach relative to their experience in other technical courses. These dimensions are grounded in Ryan and Deci's [21] self-determination theory distinction between intrinsic motivation and perceived competence, both of which are supported by instructional designs that reduce unnecessary difficulty while maintaining meaningful analytical challenge.

3. Materials and Methods

3.1 Research Design

This study employs a descriptive, one-group post-test design. Following the implementation of the worked-example-based instructional intervention throughout course MP-261 (Human Factors Engineering), students completed a structured satisfaction questionnaire. This design is appropriate for the study's primary purpose—gathering formative evidence about the instructional value of the worked-example approach—given that the assignment of a control group without access to worked examples would raise ethical concerns about equity of educational opportunity in a small-cohort graduate setting. The focus is on describing the students' experience with the intervention, not on comparing outcomes with an unexposed group. This design aligns with Kirkpatrick Level 1 (Reaction) evaluation [8], which assesses the extent to which participants find the training favorable, engaging, and relevant to their work. While Level 1 data do not directly demonstrate learning, they provide essential formative evidence for iterative improvement of instructional designs and for assessing the feasibility of scaling the worked-example approach to other courses.

3.2 Participants

Participants were 21 graduate students enrolled in course MP-261 (Human Factors Engineering) in the Graduate Program in Space Sciences and Technologies (CTE) at Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, Brazil. ITA is Brazil's premier aerospace engineering institution, and its graduate students represent a highly selected population with strong quantitative backgrounds and varying degrees of prior exposure to aviation operations. Participation in the survey was voluntary; all students who completed the course were invited to respond. The cohort was characterized by demographic and professional heterogeneity: students included recent engineering graduates with no professional experience, active-duty military aviation officers, civil aviation professionals, and researchers from adjacent engineering domains. This heterogeneity is analytically valuable, as it allows assessment of whether satisfaction with the worked-example approach varies across prior knowledge levels and professional backgrounds—a theoretically motivated question given CLT's prediction [3, 4] that worked examples are most beneficial for novice learners.

3.3 Educational Intervention

The educational intervention consisted of the complete MP-261 course, delivered over one academic semester. The course covered seven HFE analysis methodologies: Hierarchical Task Analysis (HTA) [10], Goal-Directed Task Analysis (GDTA) [12, 13], Operator-Event Sequence Diagram (OESD), Event Analysis of Systemic Teamwork (EAST), Critical Decision Method (CDM) [22], Conceptual Mapping, and Cognitive Work Analysis (CWA). For each methodology, the instructor developed a comprehensive worked example using GenAI assistance, following the pedagogical approach described by [7]. The development process followed a structured protocol: (1) the instructor identified an authentic aviation operational scenario appropriate for the method; (2) a detailed prompt was engineered specifying the method's analytical requirements, the scenario context, and the pedagogical goals; (3) the LLM generated an initial worked example; (4) the instructor reviewed the output for methodological accuracy, operational authenticity, and pedagogical clarity; and (5) the refined example was formatted into a structured template for distribution to students. Templates were distributed electronically prior to each corresponding lecture, allowing students to review the worked example before engaging with the instructor's in-class explanation. In-class sessions used the worked example as a reference point for discussion, with the instructor drawing attention to critical analytical decision points, common errors, and domain-specific nuances—consistent with Renkl's [9] recommendation that self-explanation be actively prompted rather than left to student initiative. Students then applied each method to a novel aviation scenario in an assessed assignment, without access to the worked example template during the assignment itself.

3.3.1 Hierarchical Task Analysis (HTA)

HTA is a structured ergonomic method that decomposes a complex goal into a hierarchy of sub-goals, operations, and plans [10]. Unlike a simple checklist, HTA makes explicit the logical structure governing task execution: the plans specify the sequencing conditions under which sub-goals must be performed (e.g., sequential, parallel, or conditional execution). This structural explicitness is what distinguishes HTA as an analytical tool rather than merely a descriptive one — it renders latent dependencies, coordination requirements, and cognitive load concentrations visible and amenable to systematic safety analysis. The worked example applied HTA to the Aircraft Departure Clearance to Taxi sequence of an Airbus A320, encompassing the full operational arc from the initial ATC clearance request through the final aircraft configuration for taxi commencement. The scenario involves three interacting agent groups (flight crew, ground crew, and ATC), three distinct communication channels (VHF radio, intercom, and visual signals), and multiple cockpit interfaces (overhead panel, center pedestal, ECAM, and MCDU). The HTA was presented in both graphical (indented notation) and tabular formats, the latter augmented with explicit columns for the agents involved and the tools/interfaces accessed at each operation.

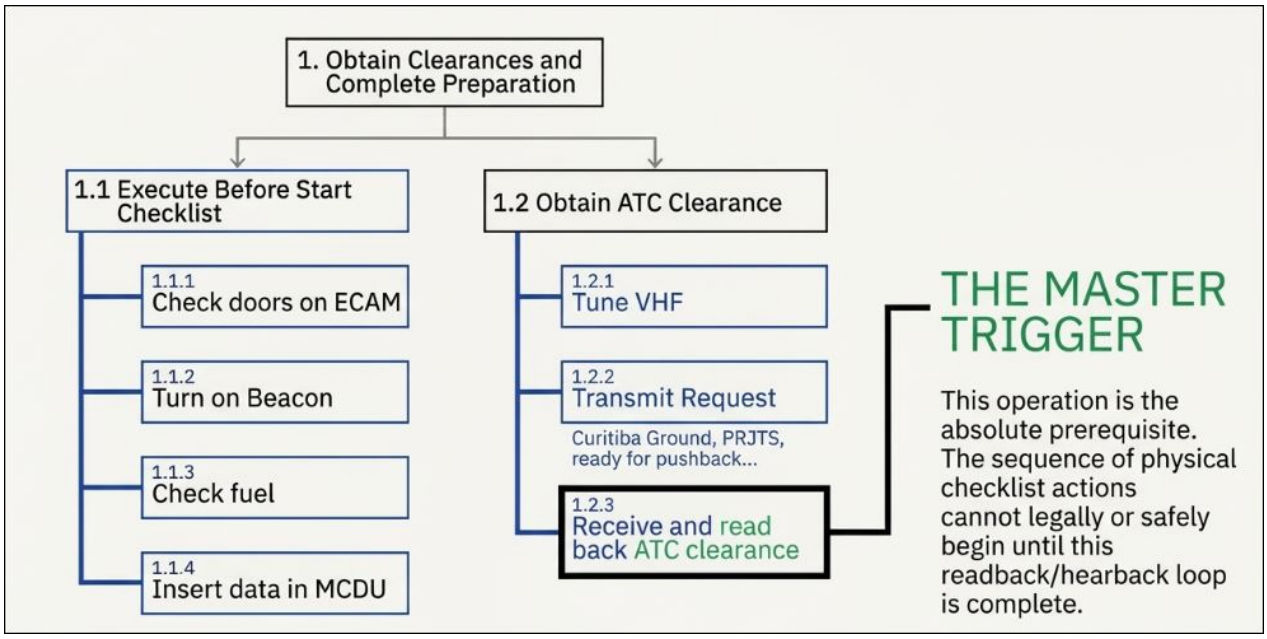


Figure 1 – Worked example using GenAI for an exerpt of the HTA - Aircraft Departure Clearance to Taxi sequence of an Airbus A320 (Source: Generated with aid of Google notebookLLM).

The example in Fig. 1 contributes to students' understanding in three specific ways. First, the decomposition of Plan 2—"Do 2.1 and 2.2 in parallel"—makes tangible the concept of concurrent cognitive load: students observe how a seemingly routine procedure forces the crew to simultaneously manage external communication with the ground crew and internal engine start monitoring, a configuration that CLT [3, 4] identifies as a high-risk loading condition. Second, the explicit mapping of operations to agents and interfaces transforms the HTA table from an abstract exercise into a distributed cognition map, allowing students to identify communication handoffs as the system's most vulnerable interfaces. Third, the analysis demonstrates HTA's function as a proactive safety tool: by walking through each plan and asking "what if" questions at critical nodes, students practise the mode of reasoning that underlies HAZOP-style procedural analysis, directly connecting HTA methodology to the SMS requirements of ICAO Annex 19 [1].

3.3.2 Critical Decision Method (CDM)

CDM is a retrospective cognitive interview technique developed by Klein, Calderwood and MacGregor to elicit the tacit knowledge underlying expert decision-making in naturalistic, high-stakes settings [22]. Structured around a set of standardised probes—covering perceptual cues, prior experience, expectations, critical information, goals, options considered, time pressure, teamwork, and system

elements—CDM is applied to a specific incident recalled by the practitioner, progressively deepening the narrative until the cognitive reasoning that drove each decision point becomes explicit. Unlike protocol analysis or think-aloud methods, CDM is particularly suited to retrospective incidents where the expert's actions cannot be reproduced in controlled conditions. The worked example presents a simulated CDM interview with an experienced Boeing 737 captain concerning an engine failure during final approach to São Paulo/Congonhas Airport in adverse weather. The nine Klein probes are applied sequentially, progressively uncovering the multimodal cue pattern that triggered situation recognition (intermittent metallic sound, yoke vibration, asymmetric yaw, and olfactory cues), the role of analogical reasoning from a prior similar event, and the trade-off logic behind choosing to continue to Congonhas rather than divert to Guarulhos or Viracopos. The example contributes to students' understanding in three specific ways. First, the probe sequence makes visible the gap between behaviour and cognition: the initial narrative describes only actions, while the probes reveal the rich perceptual and experiential substrate that produced them—illustrating why accident investigation methods relying solely on factual reconstruction systematically miss the cognitive dimension of error. Second, the parallel between the captain's "sound library" and Klein's recognition-primed decision (RPD) model provides students with a concrete anchor for the concept that experts pattern-match to prototypical situations rather than evaluate options analytically. Third, the diversion decision—where time pressure compressed the option space from three airports to one in seconds—illustrates the goal-prioritisation under degraded conditions that CDM is uniquely positioned to elicit, and that cannot be recovered from flight data recorder outputs alone.

3.3.3 Event Analysis of Systemic Teamwork (EAST)

EAST is an integrative ergonomics framework that analyses complex team performance through three interlinked network representations: a Task Network (what is done and in what order), a Social Network (who communicates with whom and via what medium), and an Information Network (what knowledge objects are shared and how they relate) [23]. The framework also incorporates a Communications Usage Diagram (CUD) critiquing the adequacy of each communication technology, a Coordination Demand Analysis (CDA) rating the teamwork intensity of each task, and an enhanced Operator-Event Sequence Diagram that integrates all three perspectives into a single timeline. By treating both human and non-human agents as nodes, EAST operationalises a distributed cognition view of safety: the system's vulnerabilities emerge not from individual errors but from the structural properties of the network itself. The worked example applies the complete EAST framework to the push-back of an Airbus A320 during the evening peak period at a high-traffic airport. The analysis covers six human agents (Captain, First Officer, Ground Controller, Tug Driver, and two Wing Walkers) and six non-human agents (aircraft systems, push-back tug, towbar, GPU, ATC, and communication infrastructure). Each of the five EAST components is developed in full, including R-based interactive visualisations of the three networks and quantitative network metrics analysis [24].

The example contributes to students' understanding in three specific ways. First, the Social Network analysis, Fig. 2, reveals a star topology centred on the Ground Controller—a structural property that quantitative centrality metrics make unambiguous—showing how a topology that is efficient under normal conditions becomes a critical single point of failure under degraded conditions, directly linking network structure to systemic risk. Second, the CDA table demonstrates that the tasks with the highest coordination demand are not the most physically demanding ones, confronting the common misconception that workload is synonymous with physical effort [4]. Third, the CUD analysis of the voice/headset versus hand-signal channel pair illustrates technological fragility: any scenario that simultaneously compromises both media—extreme noise combined with loss of line-of-sight—removes all communication redundancy, a vulnerability invisible to any single-method analysis.

3.3.4 Operator-Event Sequence Diagram (OESD)

OESD is a timeline-based graphical method that maps the sequence of tasks, decisions, communications, and system state changes performed by each operator involved in a complex multi-agent procedure [25]. Organised as a matrix in which columns represent operators (human and automated) and rows represent chronological events, OESD makes explicit the temporal coupling between agent

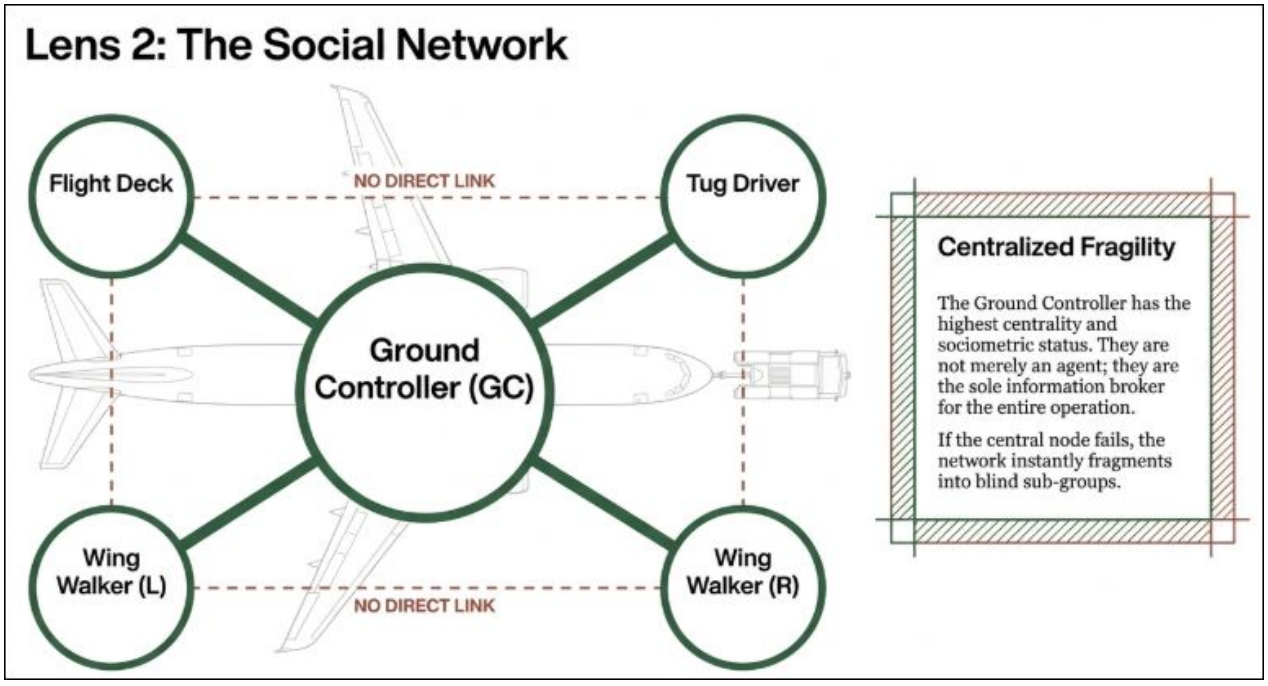


Figure 2 – Worked example using GenAI for an exerpt of the EAST - push-back of an Airbus A320 during the evening peak period at a high-traffic airport (Source: Generated with aid of Google notebookLLM).

actions—showing not only what each operator does, but when they do it relative to all other participants and to critical system state transitions. The method is widely applied in aviation, nuclear power, and defence certification contexts as formal evidence that human-system integration has been systematically analysed. The worked example applies OESD to Air-to-Air Refuelling (AAR) between a KC-130J tanker and a military helicopter, a scenario involving seven operator roles (Tanker PF and PM, Tactical Systems Operator, Crewmaster, Receiver PF and PM, and ATC) across five operational phases: Rendezvous and Formation Establishment, Pre-Contact and System Configuration, Drogue Engagement and Fuel Transfer, Disconnect and Separation, and the Breakaway emergency contingency. Each cell of the matrix is coded using a standardised symbol set distinguishing Process, Decision, Communication, Information Receipt, and System State Change events. Fig. 3 shows the inputs, processes and final representation diagram.

The example contributes to students' understanding in three specific ways. First, the comparative density of the Receiver PF and Tanker PF columns during the engagement phase makes asymmetric workload distribution immediately visible: the Receiver PF column is saturated with concurrent Process, Information, and Decision events, while the Tanker PF column shows a single sustained Process task, vividly demonstrating how platform stability is itself a human factors intervention that modulates the receiver pilot's cognitive load [3, 4]. Second, the Breakaway contingency profile illustrates the design principle of reflexive emergency procedures: the OESD shows that the entire seven-operator system must transition to a pre-programmed, non-deliberative response within seconds, underscoring why standardised phraseology and rehearsed divergent flight paths are cognitive offloading devices engineered into the system. Third, the worked example situates OESD explicitly within the certification process—showing how the diagram functions as auditable evidence for airworthiness authorities [26] that workload, role de-confliction, and human error pathways have been formally assessed, connecting the analytical method directly to the regulatory framework.

3.3.5 Goal-Directed Task Analysis (GDTA)

GDTA, proposed by Endsley [12], is a cognitive task analysis methodology designed to identify the critical goals, decisions, and information requirements of operators in complex, dynamic environments. Unlike HTA, which focuses on the what of task performance, GDTA addresses the why—the cognitive drivers behind operator actions. Its primary output is a structured mapping of the Situa-

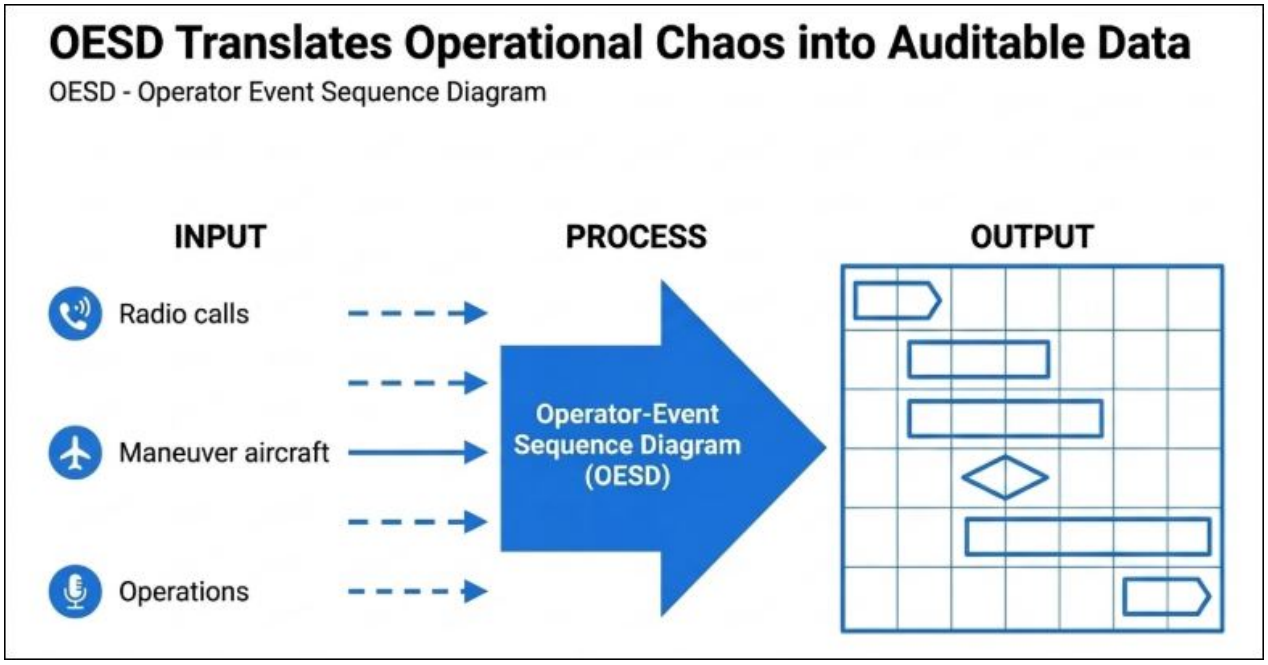


Figure 3 – Worked example using GenAI for an exerpt of the OESD - Air-to-Air Refuelling (AAR) between a KC-130J tanker and a military helicopter (Source: Generated with aid of Google notebookLLM).

tion Awareness requirements at all three levels of Endsley's SA model [12, 13]: Level 1 (Perception of raw data elements), Level 2 (Comprehension of their meaning in the context of current goals), and Level 3 (Projection of their future status). The method proceeds through knowledge elicitation from Subject Matter Experts, decomposing major goals into sub-goals, identifying the critical decision associated with each sub-goal, and then specifying the SA requirements that inform that decision. The worked example applied GDTA to the role of an Airport Tower Air Traffic Controller managing aircraft departures. The analysis covers one major goal—"Ensure Safe and Efficient Departure of Aircraft"—decomposed into four sub-goals: sequence aircraft for departure (1.1), issue takeoff clearance (1.2), monitor initial climb-out (1.3), and transfer control of aircraft (1.4). For each sub-goal, the critical decisions and the full three-level SA requirements are specified in a tabular format, supplemented by an interactive hierarchical network visualisation produced in R using the visNetwork package [24]. The example contributes to students' understanding in three specific ways. First, the juxtaposition of sub-goal 1.2 (Issue takeoff clearance) across the three SA levels makes the vertical structure of expert cognition concrete: the same runway state that appears as a raw radar return at Level 1 becomes a "sterile runway" judgement at Level 2 and a projected climb-path conflict assessment at Level 3—demonstrating that expertise consists not in perceiving more data but in constructing richer meaning from the same data. Second, the contrast between sub-goals 1.1 and 1.2 illustrates the goal-dependency of SA requirements: the information a controller needs to sequence departures efficiently (wake turbulence intervals, arrival traffic gaps) is entirely different from what is needed to issue a takeoff clearance safely (runway sterility, departure airspace), showing that SA is not a generic situational picture but a goal-relative cognitive construct [12, 13]. Third, the R-based hierarchical visualisation of the complete four-sub-goal structure gives students a concrete model of how GDTA outputs can be formatted as design inputs for interface development—connecting the analytical method directly to its application in cockpit and ATC display design, an area of active regulatory interest under ICAO human factors requirements [26].

3.3.6 Conceptual Mapping

Conceptual Mapping, as formalised by Novak and Cañas, is a knowledge representation method that organises domain concepts into a hierarchical, propositional network [27]. The method's fundamental unit is the proposition—a triplet formed by two concepts connected by a labelled linking phrase (e.g.,

"A320 Pre-Flight Check | is comprised of | Exterior Walkaround")—which encodes a single meaningful statement about the domain. Concepts are arranged hierarchically from the most general at the apex to the most specific at the leaves, while cross-links connect concepts across different branches, making non-obvious interdependencies explicit. Unlike HTA or GDTA, Conceptual Mapping does not model procedure or cognition in isolation; it models the knowledge structure of a domain, making it particularly suited to capturing the integrated mental model that an expert holds rather than the sequential steps they execute. The worked example constructed a detailed Conceptual Map of the Airbus A320 Pre-Flight Check, structured around the focus question: "What is the procedural and cognitive structure of an A320 pre-flight check performed by a pilot?" The map organises the procedure into three primary branches corresponding to the three operational phases (Preliminary Cockpit Preparation, Exterior Walkaround, and Final Cockpit Preparation), each decomposing into sub-concepts, procedures, documents, systems, and data types. The map was rendered as an interactive network in R using the visNetwork package [24], with 52 concept nodes grouped by phase and type, and 53 directed propositions encoding the procedural and logical relationships between them. A semantic deconstruction of all concepts and propositions was also provided in tabular form as a companion artefact. The example contributes to students' understanding in three specific ways. First, the cross-link between the Parking Brake concept (under Cockpit Preparation) and the Brake Wear Indicators concept (under Exterior Walkaround), labelled "enables inspection of," demonstrates the power of cross-links to surface latent procedural dependencies that are invisible in a linear checklist: a specific cockpit state is a necessary precondition for a valid physical inspection, and the map makes this causal relationship immediately apparent. Second, the identification of FMS/MCDU Programming as the node with the highest number of incoming and outgoing links—receiving route, weight, fuel, and weather data, and generating navigation and performance outputs—teaches students to use network topology as a diagnostic tool for cognitive load: high-degree nodes are simultaneously the most critical analytical hubs and the most significant error traps, a principle that applies across domains. Third, the map as a whole exemplifies the pedagogical argument underlying the entire course: the difference between knowing a checklist and understanding a procedure is precisely the difference between a linear list and a propositional network—one encodes steps, the other encodes meaning, dependencies, and safety rationale [9, 4].

3.3.7 Cognitive Work Analysis (CWA)

Cognitive Work Analysis, developed by Rasmussen and systematised by Vicente [14], is a comprehensive systems-oriented framework for analysing work in complex, sociotechnical environments. Unlike the methods described in Sections 3.3.1–3.3.6, which each focus on a specific analytical lens (task structure, expert cognition, team communication, event sequences, or knowledge representation), CWA provides an integrated, multi-phase architecture that progressively characterises the full ecological context in which human operators function. The framework is structured into five analytical phases, each addressing a distinct facet of the work system: Work Domain Analysis (WDA) models the physical and functional structure of the system in terms of purposes, functions, and physical forms using an abstraction hierarchy; Control Task Analysis (ConTA) identifies the work functions and decision requirements that arise from the domain's constraints; Strategies Analysis maps the alternative cognitive and procedural strategies operators may use to accomplish those functions; Social Organisation and Cooperation Analysis (SOCA) characterises how tasks and responsibilities are distributed across agents and roles; and Worker Competencies Analysis (WCA) specifies the cognitive and physical competencies required of operators to act within the system as modelled by the preceding phases [14, 15]. The worked examples for CWA were developed as five separate, phase-specific templates, each applied to the same aviation operational scenario to preserve analytical continuity across phases. This design choice was deliberate: a central pedagogical challenge of CWA is maintaining the systemic thread that connects the abstraction hierarchy of WDA to the competency specifications of WCA [15]. By anchoring all five templates to the same scenario, the worked examples made this vertical coherence explicit and traceable, allowing students to observe how an analytical commitment made at the WDA level propagates through each subsequent phase. The CWA template set contributed to students' understanding in three specific ways. First, the WDA abstraction hierarchy template provided the most structurally demanding cognitive exercise of the

entire course—mapping system elements simultaneously across five levels of abstraction (functional purpose, abstract function, generalised function, physical function, physical form)—and the worked example reduced the extraneous load of this multi-dimensional scaffolding task sufficiently for students to engage with the substantive analytical content [3, 10]. Second, the SOCA template made visible the distinction between who does what (HTA) and who is authorised and competent to do what under which constraints (SOCA), a nuance that is frequently collapsed by novice analysts and that the worked example made analytically explicit. Third, the five-phase structure itself communicates a meta-lesson about analytical rigour: that a complete human factors analysis of a safety-critical system cannot be reduced to any single method, but requires the layered, mutually constraining perspectives that CWA provides — a principle directly aligned with the competency demands of ICAO Annex 19 SMS frameworks [1].

3.4 Instrument

Data were collected via a structured questionnaire administered through *Google Forms* following course completion. The instrument was developed by the course instructor and comprises demographic items and ten substantive items organized into four thematic blocks. Table 1 provides an overview of the instrument structure. Items Q4 through Q6 and Q10 employ five-point Likert scales. Items Q7 and Q9 use categorical response formats. Item Q8 employs a matrix format requesting a separate five-point rating for each of seven methodology categories. Open-ended comment fields were provided after Q4, Q5, Q6, Q7, and Q10 to capture qualitative elaborations.

Item	Description	Response Format
D1	Academic background (Aeronautical Engineering / Aerospace Engineering / other engineering / other formation)	Single-choice with free-text option
D2	Current professional situation (student only / aeronautical/aerospace industry / other engineering / other)	Single-choice with free-text option
D3	Age range (20–25 / 26–30 / 31–35 / above 35 years)	Single-choice
Q4	BEFORE the course: perceived relevance of Human Factors in academic or professional projects	5-pt Likert: 1 = Never considered $\rightarrow$ 5 = Always considered
Q5	AFTER the course: extent to which structured templates facilitated the inclusion of HFE analyses in projects	5-pt Likert: 1 = Did not facilitate $\rightarrow$ 5 = Facilitated completely
Q6	Templates REDUCED mechanical complexity and allowed focus on the ESSENCE of HFE tools	5-pt Likert: 1 = Strongly disagree $\rightarrow$ 5 = Strongly agree
Q7	Most valuable template(s) for understanding course content — multiple selection permitted from 7 tools, with mandatory justification	Multiple-choice (HTA, OESD, EAST, CWA (full set), CDM, Conceptual Mapping, GDTA)
Q8	Per-methodology rating: “The template helped me understand the ESSENCE and PRACTICAL APPLICATION of this tool” — rated separately for seven tool categories	5-pt Likert per tool (1 = Did not help; 3 = Helped moderately; 5 = Helped greatly) $\times$ 7 tools
Q9	Comparison with other technical courses already taken: effect of structured templates on learning ease	5-option ordinal: Significantly easier / Easier / No significant difference / More confusing / Preferred learning without templates
Q10	Recommendation of AI-assisted structured templates for other complex engineering courses	5-pt Likert: 1 = Definitely not $\rightarrow$ 5 = Definitely yes; plus open comment field

Table 1 – Structure of the MP-261 Student Satisfaction Questionnaire.

3.5 Data Analysis Plan

Quantitative data from Likert-scale items will be analyzed using descriptive statistics (means, standard deviations, and frequency distributions). Given the ordinal nature of Likert responses and the small sample (n = 21), nonparametric statistics will be used for group comparisons: Mann-Whitney U tests for between-group differences. For Item Q8, Friedman's test was originally planned to assess whether satisfaction ratings differ significantly across methodologies. Because not all respondents provided ratings for every methodology — yielding unequal response counts per tool (n = 6–10) — the complete-block design required by Friedman's test was not achieved. The Kruskal-Wallis H test, which accommodates independent samples of unequal size, was used instead, followed by Wilcoxon signed-rank post-hoc comparisons with Bonferroni correction. All statistical analyses were conducted in R [24]. Qualitative responses from open-ended comment fields will be presented and contextualized.

4. Results

4.1 Participant Profile

Fifteen graduate students completed the post-course questionnaire, representing a response rate of 71.4% of the enrolled cohort (n = 21 enrolled; n = 15 responses). The sample was demographically heterogeneous across all three characterization dimensions, as illustrated in Fig. 4.

Regarding academic background, the majority held degrees in engineering disciplines other than aeronautical or aerospace (n = 7; 47%), with smaller groups in other formations (n = 3; 20%), aeronautical engineering (n = 3; 20%), and aerospace engineering (n = 2; 13%). Other formations specified included Law, Business Administration, Mechatronics, Mechanical Engineering, and Technology Management, reflecting the interdisciplinary character of ITA's graduate program. Professionally, most respondents were active in the aeronautical or aerospace industry (n = 7; 47%), followed by other engineering domains (n = 5; 33%), and other fields including marketing, energy, steelmaking, and artificial intelligence (n = 3; 20%). No respondent was a student-only participant, indicating that the entire sample combined graduate study with professional practice. The age distribution was skewed toward experience: 47% were above 35 years of age, 33% were aged 26–30, 13% were 31–35, and only one respondent (7%) was in the 20–25 bracket. This profile — experienced professionals with diverse disciplinary backgrounds — is consistent with the graduate cohort described in Section 3.2 and has direct interpretive relevance for the satisfaction findings.

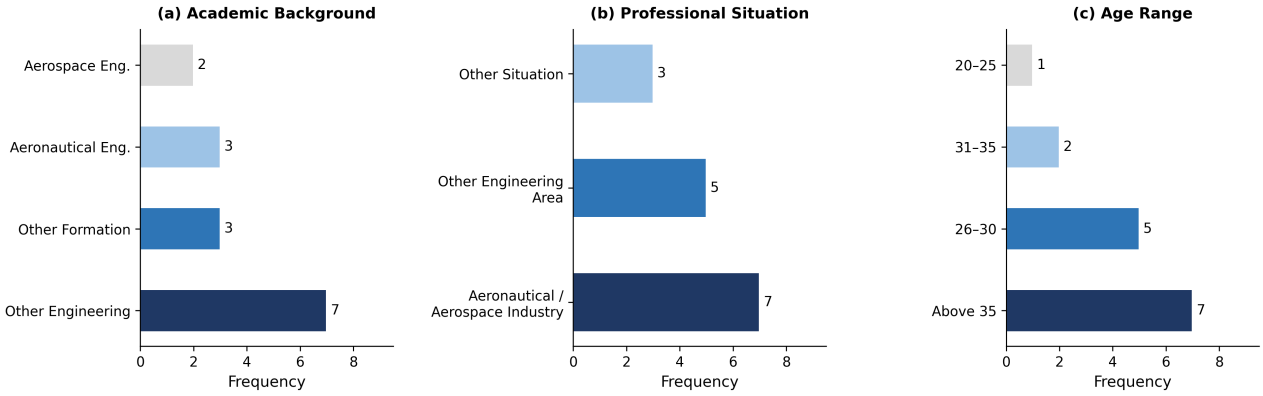


Figure 4 – Participant demographics: (a) academic background, (b) professional situation, (c) age range (n = 15).

4.2 Pre-Course Orientation toward Human Factors (Q4)

Prior to the course, students showed a moderately high orientation toward Human Factors as a relevant analytical dimension (mean = 3.93, SD = 1.16; Fig. 5 a). The distribution was notably bimodal: six respondents (40%) reported the maximum score (5 = "always considered"), four (27%) reported a score of 4, four (27%) reported a neutral score of 3, and one respondent (7%) reported a score of 1 ("never considered"). The presence of a substantial minority already highly attuned to HF is

consistent with the professional profile of the cohort — practitioners in aviation safety-critical contexts are more likely to have encountered HF requirements through regulatory frameworks such as ICAO Annex 19 prior to formal instruction. The single respondent at score 1 likely represents a participant from a domain with no prior HF exposure (e.g., marketing or law). The moderate dispersion (SD = 1.16) establishes that the cohort was not uniformly predisposed toward HF, providing meaningful variance against which post-course responses can be contextualized.

4.3 Perceived Facilitation by Structured Templates (Q5 and Q6)

Items Q5 and Q6 operationalize the primary satisfaction constructs of the study. Results were strongly positive and closely aligned across both items (Fig. 5 b,c. For Q5 ("After taking this course, did the use of structured templates facilitate the inclusion of Human Factors analyses in your projects?"), the mean rating was 4.60 (SD = 0.63), with no respondent rating below 3. Ten respondents (67%) selected the maximum score of 5 ("facilitated completely"), four (27%) selected 4, and one (7%) selected 3. For Q6 ("Did the structured templates REDUCE the mechanical complexity of the methods and allow you to focus on the ESSENCE of the HFE tools?"), the mean was 4.53 (SD = 0.64), with nine respondents (60%) at score 5, five (33%) at score 4, and one (7%) at score 3. The high means and low standard deviations for both items indicate strong and consistent positive reactions across the sample, with minimal dissent. Open-ended elaborations on Q5 and Q6 provided convergent qualitative evidence. Students described the templates as: providing a "structured and pragmatic approach" (R1); enabling a "visual organisation of processes, which is essential for studying complex concepts" (R6); functioning as "systematic guides for organising reasoning in processes, making problem situations clearer, traceable and communicable to other team members" (R7); "suppressing mechanical work and strengthening understanding of the tool's objective" (R11); and enabling a "mathematical/engineering perspective for considering human factors in research" (R11). A critical voice (R2) noted that some templates were "almost as long as reading a reference book," suggesting that template length and density remain an optimisation target. A participant with existing industry experience (R9) noted proximity to in-company templates, suggesting that the pedagogical format aligned with professional practice norms.

4.4 Most Valuable Templates (Q7)

Question 7 allowed respondents to select all templates they considered most valuable, resulting in a count of individual tool mentions across the 15 responses (Fig. 6). The results reveal a clear hierarchy of perceived value, with EAST (Event Analysis of Systemic Teamwork) emerging as the most frequently cited tool (n = 13; 87% of respondents), followed by the CWA sub-phases (SOCA: n = 10; WCA: n = 9; WDA: n = 8; ConTA: n = 7; Strategies: n = 7), HTA (n = 8; 53%), OESD (n = 6; 40%), and CDM (n = 3; 20%). Conceptual Mapping and GDTA each received only one mention.
Qualitative justifications revealed that EAST was valued for its "systemic-level analysis" and for "being very effective in analysing both individual tasks and teams, decomposing complex tasks, events, and critical interactions" (R6). The CWA sub-phases were recognised collectively by four respondents as having been "facilitated in all its stages by the template" (R4), and for "gathering multiple perspectives on the same activity, making it possible to analyse interactions and dimension workload and training requirements" (R3). HTA was cited for its clarity in "sequential process ordering" (R8) and its role in supporting thesis research (R3). OESD was highlighted for "offering a clear view of the event sequence and operator performance" (R7). The low mention counts for Conceptual Mapping and GDTA may reflect the relatively fewer respondents who rated those tools in Q8 (n = 7 and n = 6 respectively), or a perception that the templates for those methodologies were less differentiated from what could be constructed independently.

4.5 Per-Method Template Utility (Q8)

Item Q8 collected separate utility ratings for seven methodology categories. Table 2 presents the descriptive statistics, and Fig. 7 illustrates the mean ratings with standard deviations. It should be noted that the number of respondents providing ratings varied across tools (range: n = 6–10), reflecting the fact that not all respondents had applied every methodology in sufficient depth to form an evaluative judgement.

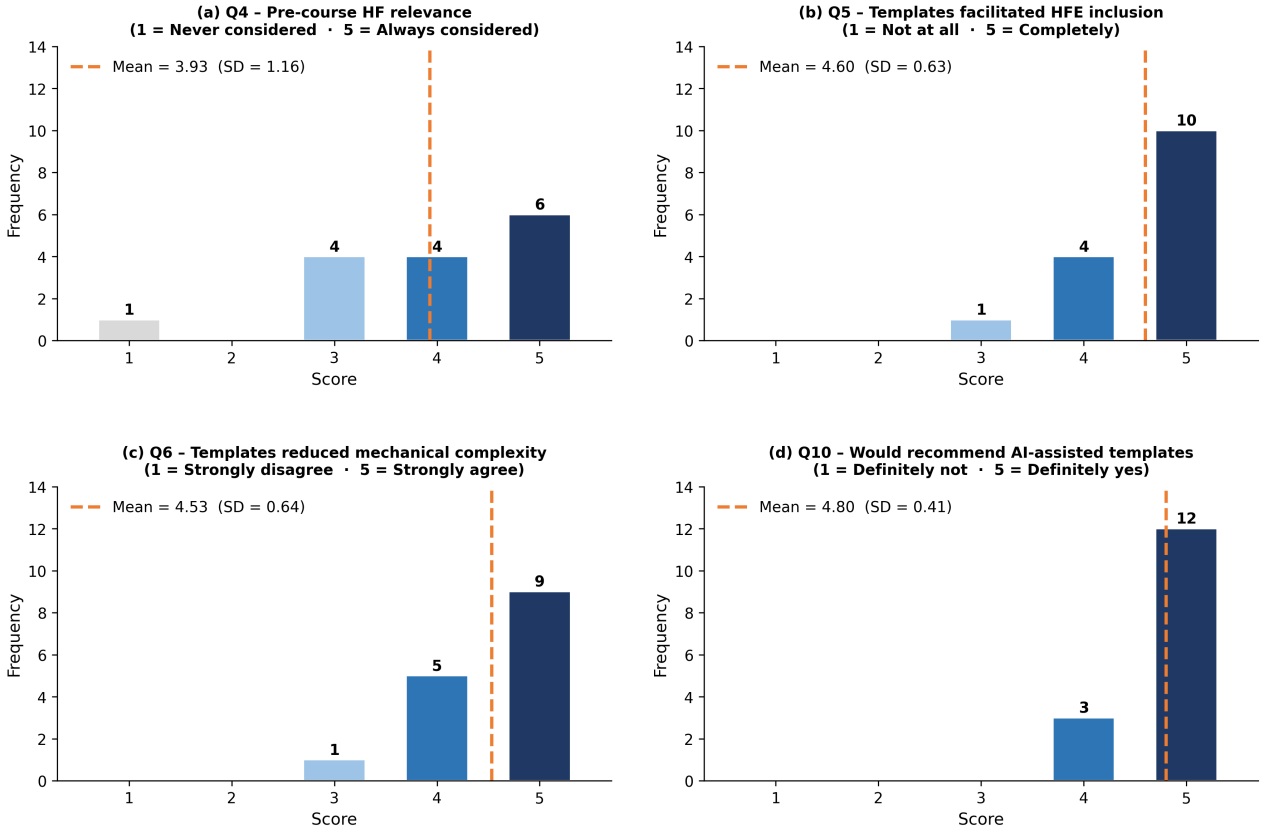


Figure 5 – Likert score distributions for items Q4, Q5, Q6, and Q10. Dashed lines indicate mean values

All mean ratings fell above the midpoint of the scale (3.0), indicating uniformly positive assessments across all methodologies. HTA and EAST achieved the highest mean rating (4.00, SD = 0.00), with perfect agreement among all raters — every respondent who evaluated these tools selected score 4. This zero-variance result warrants attention: it is possible that the Likert scale calibration influenced responses (i.e., score 4 was interpreted as "very helpful" while score 5 was reserved for "exceptional"), or that the templates for HTA and EAST were so consistently well-structured that no rater perceived a basis for differentiation. CWA received the third-highest mean (3.75, SD = 0.71), followed by OESD and CDM (both 3.67, SD = 0.82), Conceptual Mapping (3.43, SD = 0.98), and GDTA (3.33, SD = 1.03). The increasing standard deviation from HTA/EAST to GDTA suggests that the higher-order cognitive mapping tools (Conceptual Mapping, GDTA) elicited more varied reactions, consistent with CLT predictions that methods requiring more complex analytical integration benefit less uniformly from standardised templates. One respondent reported in the final comments that they were unable to submit their intended rating for CWA due to a transient issue with the Google Forms interface. This appears to have affected only that individual's entry — the majority of raters (7 of 8) successfully selected score 4 — so the episode is best read as an isolated data-entry irregularity rather than a systematic exclusion of a scale point. It is noted here for transparency but does not materially affect the reported distribution.

4.6 Comparison with Other Technical Courses (Q9)

Item Q9 asked respondents to compare the learning experience in MP-261 with other technical courses they had taken. Results were strongly positive: seven respondents (47%) selected "Made learning SIGNIFICANTLY easier" and seven (47%) selected "Made learning EASIER," yielding a combined 93% endorsing a positive comparative effect. Only one respondent (7%) selected "Made learning more confusing" — the same individual who provided the critical open comment questioning whether AI-generated content promotes genuine engagement (R1) — and no respondent selected either "Made no significant difference" or "Preferred learning without templates."

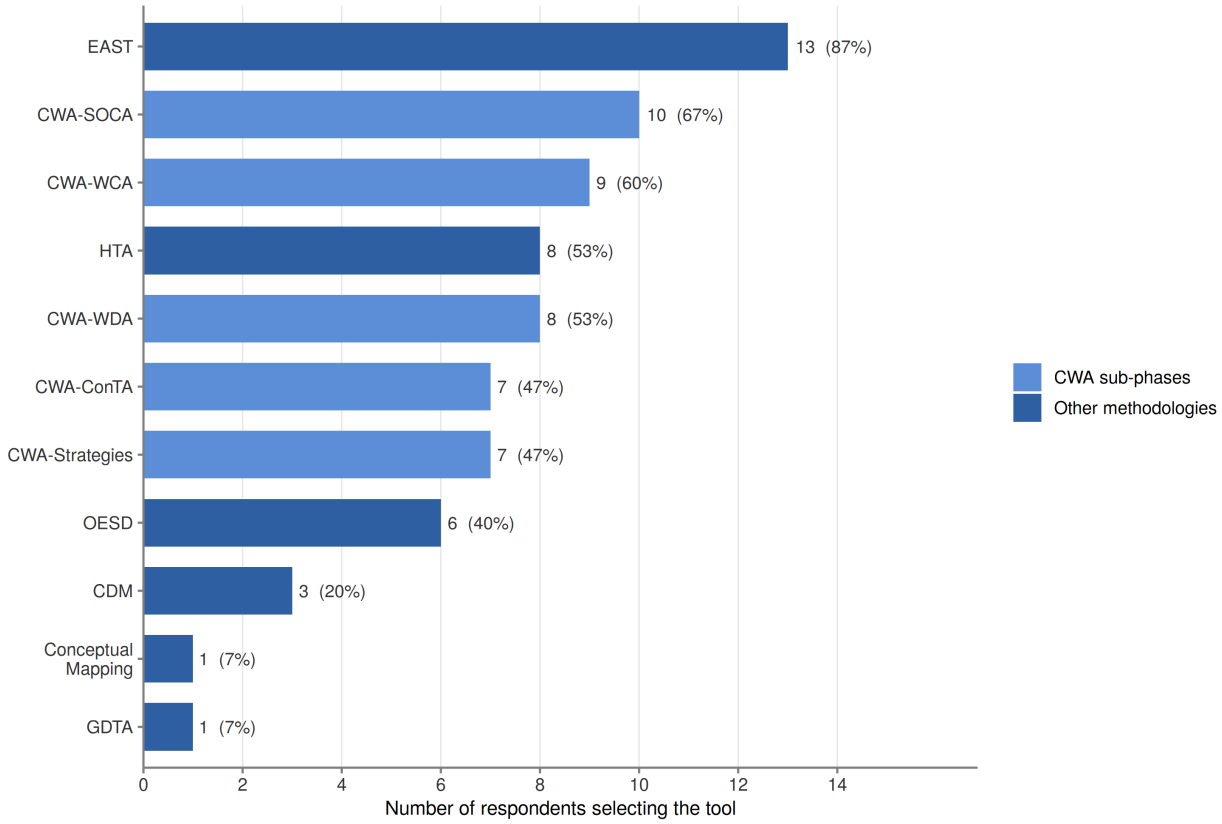


Figure 6 – Frequency of tool selection as "most valuable" (multiple selection permitted; n = 15).

Method	n	Mean	SD	Score distribution (1-5)	Rank
HTA	10	4.00	0.00	4:10 (100%)	1 =
EAST	9	4.00	0.00	4:9 (100%)	1=
CWA (full set)	8	3.75	0.71	2:1 (13%); 4:7 (87%)	3
OESD	6	3.67	0.82	2:1 (17%); 4:5 (83%)	4=
CDM	6	3.67	0.82	2:1 (17%); 4:5 (83%)	4=
Conceptual Mapping	7	3.43	0.98	2:2 (29%); 4:5 (71%)	6
GDTA	6	3.33	1.03	2:2 (33%); 4:4 (67%)	7

Table 2 – Descriptive statistics for Q8 per-methodology template utility ratings.

4.7 Recommendation for Other Engineering Courses (Q10)

Item Q10 assessed respondents' endorsement of the approach for broader application. The mean was 4.80 (SD = 0.41) — the highest of any item in the questionnaire — with 12 respondents (80%) selecting the maximum score of 5 ("Definitely yes") and three (20%) selecting score 4 (Fig. 5 d). No respondent selected a score below 4, indicating near-universal endorsement. This result is particularly noteworthy given the heterogeneity of the cohort: respondents from non-aviation domains (marketing, law, energy) rated the approach as highly recommendable as those from the aeronautical industry, suggesting that the perceived value of AI-assisted worked examples generalises beyond the specific domain of aviation HFE. Open-ended comments on Q10 revealed both enthusiasm and important caveats. One respondent (R1) articulated a principled concern: "I think the use of these templates is important; however, I still prefer traditional teaching models, as it is in this academic environment that we still have the opportunity to interact with a human to teach us. [...] The template can help as an additional tool, showing an example, but I would not use it as the central learning resource." This perspective directly invokes the pedagogical concern raised by [18]. Regarding the risk of passive consumption of AI-generated content. Another respondent (R4) flagged the importance of

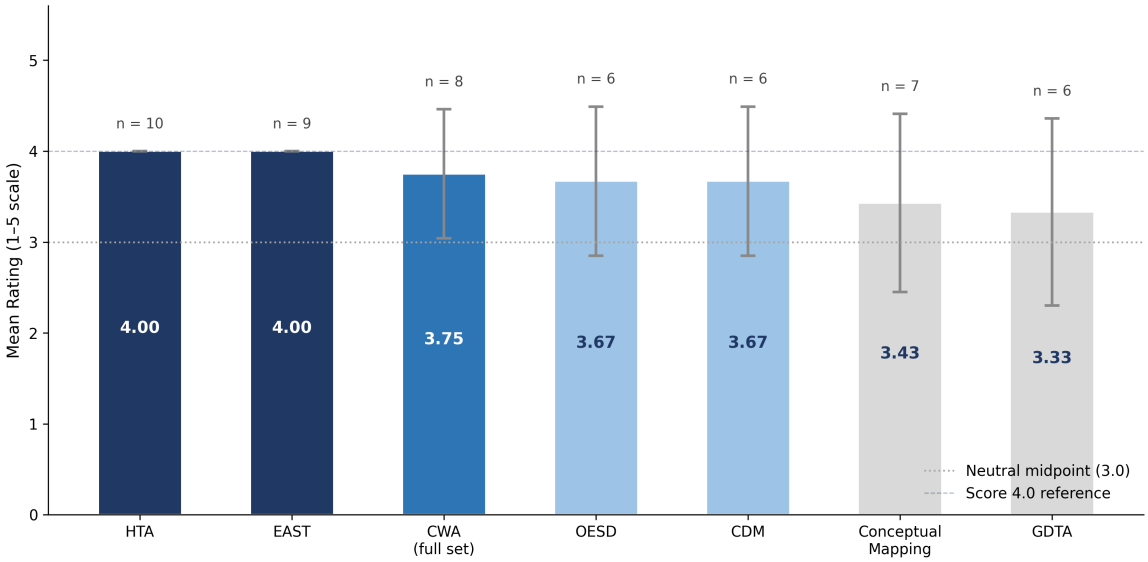


Figure 7 – Mean per-method template utility ratings (Q8). Error bars = ±1 SD; n per tool shown above bars.

"ethical use of AI" and the need for "correct prompt definitions to avoid incoherent data that, rather than reducing work, increases it due to adjustment and correction needs." A third respondent (R2) suggested that future iterations should include "cases for students to identify which methodology to apply in each scenario and how to apply it" — a call for worked examples that model methodological selection, not only methodological application. Respondent R5 recommended the inclusion of "more simulators," pointing toward immersive practice environments as a complementary learning modality.

4.8 Inferential Statistical Analysis

The analyses below implement the nonparametric tests specified in Section 3.5, with one pre-registered adaptation: Friedman's test was replaced by the Kruskal-Wallis H test for Item Q8. Friedman's test requires a complete block design in which every respondent provides a rating for every treatment condition. Because Q8 yielded unequal response counts across methodologies (range: n = 6–10), reflecting respondents' decisions not to evaluate tools they had not applied in sufficient depth, the complete-block assumption is violated. The Kruskal-Wallis H test, which accommodates independent samples of unequal size, is the appropriate nonparametric substitute. All remaining tests proceed as specified. All data vectors were reconstructed from the frequency distributions reported in Sections 4.2–4.7 and verified against reported means and standard deviations prior to analysis. All tests use continuity correction (exact = FALSE) due to ties in ordinal data. All analyses were conducted in R [24].

4.8.1 Comparison with Scale Midpoint (Items Q4, Q5, Q6, Q10)

One-sample Wilcoxon signed-rank tests assessed whether the median response on each primary Likert item differed significantly from the scale midpoint (μ = 3), representing the neutral position. Effect sizes were computed as $r = Z/\sqrt{N}$, with $r \geq 0.50$ interpreted as a large effect following Cohen's (1988) conventions. Results are presented in Table 3.

Note. V = Wilcoxon signed-rank statistic; $r = Z/\sqrt{N}$ (effect size). All tests are one-sided (H: median > 3). Q4 p-value exact: .012; Q5 exact: .0003; Q6 exact: .0004; Q10 exact: .0002. All four primary items were significantly above the scale midpoint. The satisfaction items Q5, Q6, and Q10 reached the highest significance levels (all p < .001) with very large effect sizes (r = 0.874–0.929), confirming that the positive descriptive statistics reported in Sections 4.3 and 4.7 reflect reliable population-level tendencies rather than sampling artefacts. Item Q4 — which measures pre-course orientation rather than satisfaction with the intervention — was also significantly above the midpoint (p = .012, r = 0.582), indicating that the cohort entered the course with a moderately positive prior disposition toward Human Factors, though with substantially greater inter-individual variability than the post-course

Item	n	Mean	SD	V	p	r	Magnitude
Q4	15	3.93	1.16	58.0	.0120	0.582	Large
Q5	15	4.60	0.63	105	$< .001$	0.882	Large
Q6	15	4.53	0.64	105	$< .001$	0.874	Large
Q10	15	4.80	0.41	120	$< .001$	0.929	Large

Table 3 – One-sample Wilcoxon signed-rank tests for primary satisfaction items (H_1: median > 3; one-sided).

items (SD = 1.16 vs. 0.41–0.64 for Q5, Q6, Q10). The same test applied to Q8 per-methodology ratings revealed a gradient consistent with the descriptive pattern in Section 4.5 (Table 4). HTA (p = .001) and EAST (p = .002) ratings were significantly above the midpoint; CWA reached significance at p = .021. OESD and CDM showed marginal effects (both p = .065) that did not survive the conventional threshold. Conceptual Mapping (p = .149) and GDTA (p = .242) did not differ significantly from the midpoint in the available sample, reflecting both the lower mean ratings for these methodologies and the reduced statistical power associated with their smaller response counts (n = 6–7).

Methodology	n	Mean	SD	V	p
HTA	10	4.00	0.00	55	.001
EAST	9	4.00	0.00	45	.002
CWA	8	3.75	0.71	32	.021
OESD	6	3.67	0.82	18	.065
CDM	6	3.67	0.82	18	.065
Conceptual Mapping	7	3.43	0.98	20	.149
GDTA	6	3.33	1.03	14	.242

Table 4 – One-sample Wilcoxon signed-rank tests for Q8 per-methodology ratings (H_1: median > 3; one-sided).

4.8.2 Endorsement of Comparative Learning Ease (Item Q9)

A one-sided binomial test assessed whether the proportion of respondents endorsing a positive comparative effect of structured templates (responses 'Made learning SIGNIFICANTLY easier' or 'Made learning EASIER') exceeded chance (p = 0.50). Observed: 14 of 15 respondents (93.3%) selected one of these two positive options. The test was significant (p < .001; 95% Clopper-Pearson CI: [.721, 1.000]), confirming that the proportion of respondents perceiving a positive learning facilitation effect was significantly greater than would be expected by chance. The lower bound of the confidence interval (.721) indicates that even under conservative estimation, at least 72% of the population from which this sample was drawn would be expected to endorse a positive comparative effect.

4.8.3 Differences in Utility Ratings Across Methodologies (Item Q8)

A Kruskal-Wallis H test examined whether Q8 utility ratings differed significantly across the seven methodology categories. The global test was non-significant: H(6) = 6.349, p = .385, ε^2 = 0.008. The epsilon-squared effect size of 0.008 indicates negligible explained variance in ratings attributable to methodology category after accounting for the small and unequal sample sizes per group (total N = 52, k = 7 groups, range: n = 6–10 per group). Subsequent pairwise Wilcoxon tests with Bonferroni correction yielded adjusted p-values of 1.000 for all 21 pairwise comparisons, confirming the absence of statistically detectable pairwise differences. This null result warrants careful interpretation. It does not imply that all methodologies are equally effective as targets for worked-example instruction; the descriptive difference between the highest-rated (HTA and EAST: mean = 4.00) and lowest-rated (GDTA: mean = 3.33) methodologies represents a gap of 0.67 points on a five-point scale. Rather, the non-significant Kruskal-Wallis result reflects two compounding constraints: (1) the

restricted range of observed responses in Q8 — all respondents selected either score 2 or score 4, producing a binary distribution with severely limited discriminative resolution — and (2) the small and heterogeneous n per methodology (6–10), which substantially reduces statistical power for detecting moderate between-group effects. These constraints together mean that the current dataset is underpowered for inferential comparison of methodology-specific utility, even though the descriptive pattern observed in Section 4.5 is theoretically interpretable and internally consistent with CLT predictions [3, 10]. Future studies with standardised n per methodology and a five-point response format that prompts use of all scale points would provide more resolving power for this comparison.

5. Discussion

5.1 Overall Satisfaction and CLT Alignment

The quantitative results provide consistent and statistically robust Level 1 (Reaction) evidence [20] that the GenAI-assisted worked examples approach was well-received by graduate students in MP-261. The descriptive picture reported in Sections 4.3 and 4.7 is fully corroborated by the inferential analyses in Section 4.8: one-sample Wilcoxon signed-rank tests confirmed that the median ratings for all primary satisfaction items — template facilitation (Q5: $V = 105$, $p < .001$, $r = 0.882$), complexity reduction (Q6: $V = 105$, $p < .001$, $r = 0.874$), and recommendation (Q10: $V = 120$, $p < .001$, $r = 0.929$) — were significantly and substantially above the scale midpoint of 3. All three effect sizes fall in the large range ($r > 0.50$), indicating that the positive pattern is not a statistical artefact of the small sample but reflects a reliable population-level tendency. The binomial test for Item Q9 similarly confirmed that the 93.3% endorsement rate for a positive comparative learning effect significantly exceeds what would be expected by chance (14/15; $p < .001$; 95% CI [.721, 1.000]), with the lower confidence bound indicating that, under conservative estimation, at least 72% of the population from which this cohort was drawn would be expected to report easier or significantly easier learning relative to other technical courses they had taken. Item Q4 — which measures pre-course orientation rather than satisfaction with the intervention — was also significantly above the scale midpoint ($V = 58$, $p = .012$, $r = 0.582$), but with markedly greater inter-individual variability (SD = 1.16) than the post-course items (SD range: 0.41–0.64). This dispersion is consistent with the cohort's professional heterogeneity: practitioners with prior exposure to ICAO Annex 19 Safety Management Systems [1] are more likely to have encountered formal Human Factors requirements before the course, whereas respondents from domains such as marketing or law may have had no prior contact with the discipline. The convergence of variance around the high-scoring post-course items, despite the cohort's diverse starting points, is itself a substantive finding: the worked-example intervention appears to have produced high satisfaction across a wide range of prior orientations, which extends rather than merely replicates the descriptive results of Section 4.3. These results are interpretable within the framework of Cognitive Load Theory [3, 10]. The open-ended responses directly reference the mechanism that CLT predicts: templates "suppressed mechanical work and strengthened understanding of the tool's objective" (R11); they enabled focus on "the essence of human factors rather than the concern of how best to present that information" (R7); they allowed the "step-by-step" logic of each method to be internalised without the simultaneous burden of format construction (R4). These testimonies map closely onto the reduction of extraneous load that [9] identifies as the primary mechanism through which worked examples facilitate schema formation. The statistical confirmation — with large-magnitude effects for all three satisfaction constructs — provides a level of evidence that goes beyond the descriptive pattern and validates the claim that the approach was not merely received positively but perceived as fundamentally useful in the specific cognitive terms that CLT predicts. This finding is consistent with the engineering education literature showing that AI-generated worked examples are rated as more helpful than traditional textbook examples when they provide explicit step-by-step reasoning [19], and lends quantitative weight to the satisfaction signals observed by [19] in undergraduate engineering populations. A complementary finding warrants dedicated analytical attention: the convergence of satisfaction ratings across a deliberately heterogeneous cohort. Respondents included military aviation officers with direct operational experience in SMS frameworks, engineers from adjacent domains with no prior HFE exposure, and professionals from entirely distinct fields — law, marketing, steelmaking, and artificial intelligence. This breadth of

backgrounds represents precisely the kind of variance that CLT predicts should moderate the effectiveness of worked examples: the effect is theoretically strongest for novice learners and potentially redundant for experts with well-consolidated schemas [3, 4]. Yet Q5, Q6, and Q10 ratings exhibit low dispersion (SD = 0.41–0.64) and an absence of responses below the scale midpoint regardless of declared background — a pattern suggesting that the approach is robust to substantial variation in prior knowledge. This robustness carries implications that extend well beyond the academic setting: HFE training programmes in the aeronautical industry routinely bring together, in the same cohort, design engineers, flight crew, air traffic controllers, and safety managers — precisely the heterogeneity reproduced in the present sample. If GenAI-assisted worked examples produce consistently high satisfaction and perceived utility across this mixed profile, this constitutes preliminary evidence that the strategy can be scaled to organisational training contexts without requiring differentiation by background level, substantially simplifying the design and logistics of SMS capacity-building programmes. Future studies incorporating subgroup analyses by prior formation and professional experience will allow this hypothesis to be tested with adequate statistical power.

5.2 Differential Effectiveness Across Methodologies

The per-methodology ratings (Q8) and mention frequencies (Q7) reveal an interpretable descriptive pattern of differential effectiveness, but the inferential picture is more nuanced and constrains the conclusions that can be drawn from the descriptive data alone. At the descriptive level, HTA and EAST achieved the highest mean utility ratings (4.00, SD = 0.00) in Q8, and ranked highest in Q7 by mention frequency (87% and 53% of respondents, respectively). These two methodologies share a structural property that is directly relevant to CLT: they produce highly visual, spatially representable output formats — HTA's hierarchical decomposition tree and EAST's integrated network-of-networks — that are amenable to worked examples integrating text and diagrams, an instructional configuration with well-documented advantages under dual-coding theory [4]. CWA sub-phases were cited most frequently as most valuable in Q7 (SOCA: 67%; WCA: 60%; WDA: 53%), and received the third-highest Q8 mean rating (3.75, SD = 0.71), consistent with CLT's prediction that worked examples benefit most prominently for high-intrinsic-load, multi-step tasks [3, 12]. Conceptual Mapping and GDTA received the lowest Q8 ratings (3.43 and 3.33) and the highest standard deviations (0.98 and 1.03), as well as the fewest mentions in Q7 (one each), suggesting that both the effectiveness and the perceived value of worked examples for these methodologies were more variable across the cohort. The one-sample Wilcoxon tests in Section 4.8 provide a methodologically precise complement to this descriptive gradient. HTA ($p = .001$) and EAST ($p = .002$) ratings were reliably above the scale midpoint; CWA reached the conventional threshold ($p = .021$). OESD and CDM showed marginal effects (both $p = .065$) that did not survive the conventional alpha of .05. Conceptual Mapping ($p = .149$) and GDTA ($p = .242$) did not differ significantly from the midpoint in the available sample. This gradient parallels the descriptive ordering — methodologies with higher mean ratings also more consistently exceed the midpoint — and provides a finer-grained characterisation of where the evidence for utility is statistically supported versus merely suggestive. However, the global Kruskal-Wallis test on Q8 ratings across the seven methodology categories was non-significant: $H(6) = 6.349$, $p = .385$, $\varepsilon^2 = 0.008$. Subsequent pairwise Wilcoxon comparisons with Bonferroni correction yielded adjusted p-values of 1.000 for all 21 pairs. This null result must be interpreted carefully. It does not imply that the descriptive differences between methodologies are meaningless or that all tools benefit equally from worked examples; the distance between HTA/EAST (mean = 4.00) and GDTA (mean = 3.33) represents a gap of 0.67 scale points that is theoretically coherent and internally consistent with CLT. Rather, the non-significant Kruskal-Wallis result reflects two compounding statistical constraints. First, Q8 responses exhibited a severely restricted range: across all 52 individual ratings, respondents selected only scores 2 or 4, producing a binary distribution that forfeits most of the discriminative resolution available in the five-point scale. This restriction may reflect a common response set — in which score 4 was interpreted as the normative positive response and score 2 as the only alternative — or a genuine ceiling effect in which respondents found it difficult to meaningfully differentiate between methodologies at the upper end of the scale. Second, the unequal and small group sizes per methodology (n = 6–10) substantially reduce the power available for between-group detection, particularly when the response distribution is bimodal. Together, these constraints mean that the current dataset cannot statistically

resolve the descriptive differences that CLT predicts — and that the design of future iterations of Q8 should attend explicitly to response format and scale calibration. Providing anchored descriptors for each scale point (e.g., score 5 explicitly defined as "exceeded my expectations") and establishing a target minimum n of 15 per methodology would substantially increase the resolving power of this analysis in subsequent studies. These considerations have a direct implication for how Section 5.2's conclusions should be communicated: the descriptive ordering of methodology-specific effectiveness is real and theoretically interpretable, and the one-sample tests provide item-level statistical support for the three highest-rated methodologies. What the evidence does not yet support is a claim that the differences between methodologies are statistically distinguishable from one another. Future work should treat the present pattern as a hypothesis-generating pilot rather than a confirmed ranking.

5.3 Critical Voices and Design Implications

The most substantive critical response (R1) raises a pedagogically important concern: that the circular use of AI to generate templates and AI to complete assignments may shortcut the deep engagement necessary for genuine learning. This concern resonates with [18] warning that LLM-generated content risks promoting passive content consumption rather than active schema construction. It also aligns with the broader critique of worked examples research: examples are most effective when learners actively self-explain the presented solutions [12], and least effective when they are used as copy-paste starting points. The statistical confirmation that satisfaction with complexity reduction (Q6) and facilitation (Q5) is robust — both with very large effect sizes — does not resolve this concern; high perceived utility and active cognitive engagement are not synonymous. Indeed, the risk identified by R1 is precisely that templates may reduce extraneous load so effectively that they also reduce the germane load — the effortful processing that drives schema formation [3, 10]. The instructional design response is not to abandon worked examples but to engineer the conditions for active engagement — through self-explanation prompts, completion problems (partially worked examples with blanks at critical decision points), and, as suggested by respondent R2, deliberate methodological selection tasks that require students to judge which analytical framework is appropriate before applying it. These design enhancements align with the "fading" strategy advocated by [9] and with the "AI as teammate" role described by [7], in which the AI generates the first pass and the student must critically evaluate, extend, or refute it. The concern about prompt quality and correction overhead (R4) is practically significant for instructors considering adoption. The development protocol described in Section 3.3 — which includes instructor review for methodological accuracy, operational authenticity, and pedagogical clarity — is specifically designed to catch AI-generated inaccuracies before distribution. However, the respondent's observation suggests that students may also encounter AI-generated content in their own use of LLMs for assignment completion, where no such quality filter exists. This points to a need for explicit instruction in critical evaluation of AI outputs as a meta-competency in courses that integrate GenAI tools. The finding that recommendation rates were near-universal (Q10: mean = 4.80, $p < .001$, $r = 0.929$) indicates that students broadly endorse the approach despite being aware of this risk — suggesting that any future redesign should amplify the quality-filtering mechanisms rather than withdraw the worked-example strategy altogether. The proposal by respondent R2 for methodological selection tasks — exercises requiring students to identify which tool to apply before applying it — deserves specific attention in relation to the differential effectiveness pattern discussed in Section 5.2. The methodologies for which the worked-example effect was weakest (Conceptual Mapping and GDTA, both with non-significant one-sample tests and the highest rating variability) are precisely those requiring the most open-ended knowledge construction, where a single worked example is least transferable as a template. For these methodologies, a sequenced curriculum using multiple worked examples at varying complexity levels, combined with explicit selection tasks that contrast Conceptual Mapping with HTA (both model the same procedure from different epistemic angles), may yield more consistent utility ratings than the single-example approach used in the present iteration.

5.4 Limitations

Several limitations constrain the interpretation of these findings. First, the sample size (n = 15) is small and drawn from a single course at a single institution, limiting statistical power and generalis-

ability. While the one-sample Wilcoxon tests and binomial test yielded consistent and large effects for the primary satisfaction items, the power available for the between-methodology comparisons in Q8 was substantially lower due to the combination of small group sizes (n = 6–10 per methodology) and the restricted binary response distribution (only scores 2 and 4 observed), which together preclude reliable detection of the differences that the descriptive pattern suggests. Future studies should establish a minimum target of n = 15 evaluations per methodology and revise the Q8 scale calibration to encourage use of the full five-point range. Second, the one-group post-test design does not permit causal inference: the observed satisfaction cannot be attributed exclusively to the worked examples, as other features of the course — instructor quality, scenario selection, and the novelty of the GenAI framing itself — may have contributed independently. A quasi-experimental design comparing worked-example and non-worked-example sections of the same course would be required to isolate the instructional effect. Third, the Q8 item yielded variable response rates across tools (n = 6–10), creating unequal precision across methodology comparisons and precluding the use of Friedman's test as originally specified in Section 3.5. The Kruskal-Wallis H test, used as the appropriate substitute for unequal-n independent samples, returned a non-significant global result ($H(6) = 6.349$, $p = .385$, $\varepsilon^2 = 0.008$), which may reflect genuine absence of differential effectiveness, insufficient statistical power, or the scale-calibration artefact noted above — the current data cannot distinguish between these explanations. Fourth, the Likert scale appears to have been interpreted with restricted range in Q8 — no respondent selected scores 1, 3, or 5 for any methodology. This may reflect a common response set anchored at score 4 as the normative positive response, a ceiling effect for the highest-rated items, or a genuine perception that the distinction between scores is not meaningful given the brevity of each methodology's evaluation period. Regardless of its source, this restriction substantially reduces the discriminative sensitivity of Q8 and limits both the descriptive and inferential conclusions that can be drawn from it. Fifth, one respondent reported difficulty submitting their intended rating for the CWA item due to a transient issue with the Google Forms interface (Section 4.5). This appears to have affected a single entry rather than reflecting a systematic exclusion of a scale point: the great majority of CWA raters (7 of 8; 87%) successfully selected score 4, as shown in Table 2. The episode is disclosed here for transparency but is unlikely to have materially biased the CWA mean. Sixth, a significant limitation concerns the dual role of the instructor, who both designed and delivered the worked-example intervention and independently designed and administered the satisfaction questionnaire. This creates a substantive risk of self-assessment bias: students evaluating a course while enrolled in an ongoing academic relationship with — and subject to grading by — the same individual who authored the instructional materials may be susceptible to social-desirability effects that inflate reported satisfaction independent of the intervention's actual merits. No independent or blinded party was involved in instrument design, administration, or data collection, and response anonymity relative to the instructor cannot be fully guaranteed in a small-cohort setting. This conflict of interest should be weighed when interpreting the uniformly positive results reported in Sections 4.3–4.7, and future studies should employ third-party survey design and administration, with responses collected and anonymised independently of the instructor, to mitigate this source of bias.

Finally, these are Level 1 (Reaction) data: while utility reactions are significantly correlated with learning transfer [20], they do not directly measure learning outcomes (Level 2), behaviour change (Level 3), or organisational impact (Level 4), which would require longitudinal follow-up assessment. The statistically large effect sizes for Q5, Q6, and Q10 establish that the positive reactions are genuine and consistent, but they do not resolve the question — raised pointedly by respondent R1 — of whether perceived utility translates into deeper schema formation and durable analytical competence. Demonstrating this translation is the essential next step for this research programme.

6. Conclusions

This study evaluated student satisfaction with a GenAI-assisted worked examples approach in a graduate-level Human Factors Engineering course at ITA, across 15 respondents representing a heterogeneous cohort of experienced engineering professionals. Inferential analyses confirmed that all primary satisfaction items were significantly and substantially above the scale midpoint: templates were perceived as facilitating HFE inclusion in projects (Q5: mean = 4.60, $p < .001$, $r = 0.882$), reduc-

ing the mechanical complexity of the methods (Q6: mean = 4.53, $p < .001$, $r = 0.874$), and improving learning ease relative to other technical courses in the respondents' academic trajectory (Q9: 93.3% positive endorsement, $p < .001$). The recommendation rate was the highest of any measured item (Q10: mean = 4.80, $p < .001$, $r = 0.929$), with large effect sizes across all constructs confirming that these positive reactions reflect a reliable population-level tendency rather than a sampling artefact. At the methodology level, one-sample tests revealed a statistically supported gradient: HTA and EAST ratings were reliably above the midpoint ($p < .01$), CWA reached significance ($p = .021$), while Conceptual Mapping and GDTA did not — though the global Kruskal-Wallis test was non-significant [$H(6) = 6.349$, $p = .385$, $\varepsilon^2 = 0.008$], indicating that the current dataset, constrained by small per-methodology sample sizes and a restricted binary response distribution in Q8, lacks sufficient power to statistically distinguish between methodologies. The descriptive ordering is theoretically coherent and consistent with CLT predictions [3, 4, 9], but should be treated as hypothesis-generating rather than confirmatory. These findings provide statistically grounded, Level 1 evidence [20] supporting the integration of GenAI-developed worked examples as a well-received and pedagogically promising strategy for teaching complex HFE analysis methodologies in graduate engineering education. The critical voices within the cohort highlight two design parameters that must be addressed in future iterations: the risk that well-structured templates reduce germane load alongside extraneous load — shortcutting the active self-explanation that drives schema formation — and the dependence of template quality on carefully engineered prompts and instructor review. Both challenges are addressable through progressive fading strategies, self-explanation prompts, methodological selection tasks, and explicit instruction in critical AI output evaluation [9, 18]. Future work should prioritise Level 2 evaluation of learning outcomes, extension to multiple cohorts and institutions to determine whether the strong satisfaction signal observed here translates into durable analytical competence in safety-critical engineering practice.

7. Generative AI Disclosure

During the preparation of this work, the authors used Generative AI to improve the language and readability of the paper and to create the visual figures. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the final version of the manuscript.

8. Acknowledgement

This study was financed in part by the Coordenacao de Aperfeicoamento de Pessoal de Nıvel Superior– Brasil (CAPES)– Finance Code 001.

9. Contact Author Email Address

The contact author email address should appear explicitly in this section to facilitate future contacts. For example, mailto: moacyr@ita.br

10. Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

References

[1] ICAO J. *Annex 19 to the Convention on International Civil Aviation: Safety Management*. International Civil Aviation Organization., 2nd edition, 2018.

[2] Neville A. Stanton, Paul M. Salmon, Laura A. Rafferty, Guy H. Walker, Chris Baber, and Daniel P. Jenkins. *Human Factors Methods: A Practical Guide for Engineering and Design*. Ashgate Publishing, 2nd edition, 2013.

[3] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

[4] John Sweller, Jeroen J. G. van Merriënboer, and Fred Paas. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31:261–292, 2019.

[5] Robert K. Atkinson, Sharon J. Derry, Alexander Renkl, and Donald Wortham. Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, 70(2):181–214, 2000.

[6] Alexander Renkl. Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21(1):1–29, 1997.

[7] Ethan R. Mollick and Lilach Mollick. Assigning AI: Seven approaches for students, with prompts. Technical report, The Wharton School, University of Pennsylvania, 2023. SSRN Working Paper. `https://ssrn.com/abstract=4475995`.

[8] Donald L. Kirkpatrick and James D. Kirkpatrick. *Evaluating Training Programs: The Four Levels*. Berrett-Koehler Publishers, San Francisco, CA, 3rd edition, 2006.

[9] Alexander Renkl. Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, 38(1):1–37, 2014.

[10] John Annett and Keith D. Duncan. *Task Analysis and Training Design*. Number 6 in Training Information Paper. H.M.S.O., London, UK, 1967. Department of Employment and Productivity.

[11] Moacyr M. Cardoso-Júnior, H. C. Marques, and M. V. de Lima. A framework proposal for augmented reality considering human errors in sequential tasks. In S. Jin, J. H. Kim, Y. K. Kong, J. Park, and M. H. Yun, editors, *Proceedings of the 22nd Congress of the International Ergonomics Association*, volume 56 of *Springer Series in Design and Innovation*, Singapore, 2025. Springer.

[12] Mica R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64, 1995.

[13] Mica R. Endsley and Debra G. Jones. *Designing for Situation Awareness: An Approach to User-Centered Design*. CRC Press, Boca Raton, FL, 2nd edition, 2012.

[14] Kim J. Vicente. *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. Lawrence Erlbaum Associates, Mahwah, NJ, 1999.

[15] Catherine M. Burns and Judi R. Hajdukiewicz. *Ecological Interface Design*. CRC Press, Boca Raton, FL, 2004.

[16] Steven T. Shorrock and Claire A. Williams, editors. *Human Factors and Ergonomics in Practice: Improving System Performance and Human Well-Being in the Real World*. CRC Press, Boca Raton, FL, 2016.

[17] Christopher D. Wickens, Justin G. Hollands, Simon Banbury, and Raja Parasuraman. *Engineering Psychology and Human Performance*. Pearson, Upper Saddle River, NJ, 4th edition, 2013.

[18] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Andreas Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Wendler, and Gjergji Kasneci. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[19] Chih-Pu Dai, Fengfeng Ke, and Yi Pan. Investigating AI-generated explanation feedback for supporting college students' learning of complex concepts. *Computers & Education*, 203:104820, 2023.

[20] George M. Alliger, Scott I. Tannenbaum, Winston Bennett, Holly Traver, and Abigail Shotland. A meta-analysis of the relations among training criteria. *Personnel Psychology*, 50(2):341–358, 1997.

[21] Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000.

[22] Gary A. Klein, Roberta Calderwood, and Donald MacGregor. Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(3):462–472, 1989.

[23] Neville A. Stanton, Chris Baber, and Don Harris. *Modelling Command and Control: Event Analysis of Systemic Teamwork*. Ashgate Publishing, Aldershot, UK, 2008.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[25] Paul M. Salmon, Neville A. Stanton, Guy H. Walker, Chris Baber, Daniel P. Jenkins, Richard McMaster, and Mark S. Young. What really is going on? review of situation awareness models for individuals and teams. *Theoretical Issues in Ergonomics Science*, 9(4):297–323, 2008.

[26] International Civil Aviation Organization. Annex 19 to the convention on international civil aviation: Safety management. Technical report, International Civil Aviation Organization, Montréal, Canada, 2018.

[27] Joseph D. Novak and Alberto J. Cañas. The theory underlying concept maps and how to construct and use them. Technical Report IHMC CmapTools 2006-01 Rev 01-2008, Florida Institute for Human and

Machine Cognition, Pensacola, FL, 2008.