Framework for Multimodal Cognitive Load Analysis in Safety-Critical Systems: An ATC Simulation Case Study

Jonas Pöhler
University of Siegen
Siegen, Germany
jonas.poehler@uni-siegen.de

Antonia Vitt
University of Siegen
Siegen, Germany
antonia.vitt@student.uni-siegen.de

Nadine Flegel
Trier University of Applied Sciences
Trier, Germany
n.flegel@inf.hochschule-trier.de

Tilo Mentler
Trier University of Applied Sciences
Trier, Germany
mentler@hochschule-trier.de

Kristof Van Laerhoven
University of Siegen
Siegen, Germany
kvl@eti.uni-siegen.de

Abstract

Controlled studies in safety-critical domains such as Air Traffic Control (ATC) are inherently difficult, making high-fidelity simulators essential for research. However, existing simulation environments are often complex and expensive facilities that are only available at selected locations (e.g. flight simulators) or lack necessary realism, limiting their use in Human-Computer Interaction (HCI) research. This paper presents a framework that addresses this gap, demonstrating how a more realistic, sensor-enhanced simulation environment can be developed in a comparatively low-cost manner. Following the Design Science Research (DSR) methodology, we integrated the open-source BlueSky ATC engine with a custom frontend and multiple sensor modalities (e.g., eye-tracking, PPG, respiration). Our preliminary evaluation in a landing scenario case study confirms the framework's effectiveness in capturing rich physiological and behavioral data corresponding to cognitive load. We present the system architecture, assess the DSR process, and release the framework as an open source tool to foster further research.

CCS Concepts

• **Human-centered computing** → **User interface programming**; **Laboratory experiments**; Empirical studies in HCI.

Keywords

Cognitive Load, Air Traffic Control, Human-Computer Interaction, Design Science Research, Multimodal Sensing, Safety-Critical Systems

1 Introduction

Research involving air traffic controllers (ATC) presents significant challenges due to the high complexity and safety-critical nature of the domain, which limits direct access for controlled studies. Simulators offer a viable alternative, but their utility is contingent on a high degree of realism. While sophisticated solutions exist, they are often proprietary, expensive, or lack the customization capabilities needed for detailed human-computer interaction (HCI) research. Recognizing this gap, we demonstrate how a highly realistic and customizable ATC simulation environment can be developed using a cost-effective, open-source approach. Our system utilizes the BlueSky ATC simulation engine as a backend, integrated with a bespoke frontend and a suite of physiological sensors designed specifically for comprehensive user research. This frontend facilitates detailed tracking and analysis of user interactions, enabling the collection of rich behavioral and physiological data critical for evaluating cognitive load and decision-making processes. To comprehensively assess cognitive load, our system interfaces with multiple sensor modalities including respiration tracking, photoplethysmography (PPG), thermal imaging, and eye-tracking technologies.

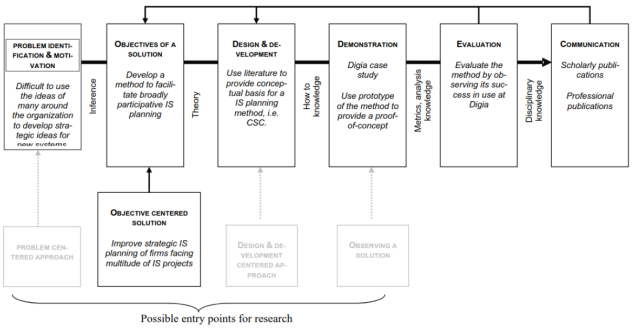

Figure 1: Research Process from Peffers et al. [12]

The development of our system employed the structured Design Science Research (DSR) methodology as proposed by Peffers et al. [12]. An additional objective of this paper is to critically assess whether the DSR process is effective for developing applications in broader safety-critical interaction contexts.

To foster collaboration and facilitate further research in the safety-critical HCI domain, the complete system, including the customized frontend, sensor integration scripts, and analysis tools, will be made openly available to the research community.

Mensch und Computer 2025 – Workshopband, Gesellschaft für Informatik e.V., 31. August - 03. September 2025, Chemnitz, Germany

https://doi.org/10.18420/muc2025-mci-ws01-191

2 Related Work

Design Science Research (DSR) has become an influential methodology in information systems (IS) research, emphasizing the creation and evaluation of innovative artifacts to address complex real-world problems [5, 7]. The DSR methodology typically follows a systematic process consisting of problem identification, solution design, development, demonstration, evaluation, and communication [13]. This structured approach aims to produce both descriptive and prescriptive knowledge, effectively balancing research rigor and practical relevance [8].

DSR has been successfully applied in various domains, including aviation training [2], product development [6], project scheduling [16], and ontology-based applications [10]. In recent years, the integration of DSR with data science has further expanded its utility, particularly in healthcare settings, by enabling innovative solutions that enhance human and social capabilities [4, 9].

Despite its growing prominence, some challenges remain regarding clarity around DSR's core concepts, potentially limiting its broader acceptance and impact [1]. Efforts to refine frameworks and design theories continue to enhance the rigor and comprehensibility of DSR practices [3, 18]. Nonetheless, the versatility and pragmatic focus of DSR make it particularly valuable for investigating emerging technologies and advancing established practices, even in contexts with limited existing theoretical foundations [17].

Given the complexity and critical nature of ATC research, adopting the DSR framework provides a structured approach to develop, evaluate and refine innovative simulation environments, ensuring that they meet the demands of both rigor of research and practical applicability. Complementing this methodological foundation, existing ATC simulators provide important context for evaluating our own approach. Simulators such as LABY and MACS have been used to explore user interaction paradigms and support high-complexity safety-critical scenarios. LABY incorporates input modalities such as mouse, stylus, and eye-tracking to facilitate interface testing and performance benchmarking in conflict resolution tasks [11]. MACS, a multi-role platform developed for distributed simulation, supports detailed safety assessments within advanced scenarios like Metroplex operations and converging runways [14].

Despite these capabilities, many simulators lack sufficient support for scenario customization and integrated physiological monitoring, which limits their suitability for nuanced human-computer interaction studies in safety-critical domains. Our work addresses this gap by extending simulation capabilities with sensor integration and a focus on replicating realistic workloads for user evaluation.

3 Methodology

The development of our framework followed the structured Design Science Research (DSR) methodology, which emphasizes the creation and evaluation of innovative artifacts to solve real-world problems. Our process was guided by iterative cycles of problem identification, solution design, and evaluation. The primary objective was to develop a robust framework capable of evaluating cognitive load within control room environments, initially implemented for emergency control rooms and now extended to ATC scenarios.

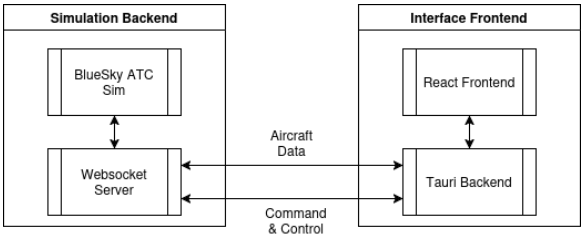

Figure 2: System architecture of the simulation framework, illustrating the separation between the Simulation Backend (powered by BlueSky) and the Interface Frontend (built with React and Tauri). Communication for aircraft data and control commands is handled via a WebSocket interface

As a key step in the DSR process, we engaged in detailed expert consultations to ensure relevance and realism. We modeled our simulation on real-world ATC centers, particularly the training facilities of the Deutsche Flugsicherung (DFS). Through structured interviews with an active air traffic controller, we iteratively refined the interface design to reflect operational workflows and ensure authenticity.

However, practical considerations led us to adjust certain interaction modalities for enhanced usability and clarity. Unlike traditional ATC operations, which rely heavily on voice communication, our simulation incorporates an interactive graphical user interface (GUI) to manage aircraft movements. This design decision was a deliberate trade-off. To achieve the high-precision, time-stamped data logging of every user command required for our multimodal sensor analysis, a GUI was essential. We acknowledge that this substitution fundamentally alters the task's nature and the operator's cognitive load profile compared to real-world operations. Despite this modification, the fundamental challenge of safely coordinating and separating aircraft in a dynamic environment remains intact. These design decisions were made in close collaboration with our consultant to balance authenticity with the needs of our research objectives.

4 Implementation and Demonstration

Our simulation framework is built upon the BlueSky Open Air Traffic Simulator, an open-source project developed at TU Delft. The primary goal of BlueSky is to advance Air Traffic Management (ATM) research by providing tools and data for simulating air traffic environments that are openly shareable. Thereby making research results more comparable and repeatable. It was designed as a fully open-source and open-data simulation tool that is multi-platform, easy to use, and free. A key objective is to provide a user-friendly, high-fidelity simulator that does not require users to have deep computer science knowledge; for example, it uses simple, editable text files for scenario definitions. The simulator is written in Python, which supports its multi-platform nature and makes it accessible to a large academic community. BlueSky employs a simple scenario language called TrafScript to facilitate ease of use for new users. Unlike server-based approaches like CASSIOPEIA, BlueSky is designed to be fully downloadable, allowing researchers to run it on a local machine and modify the source code to fit their specific

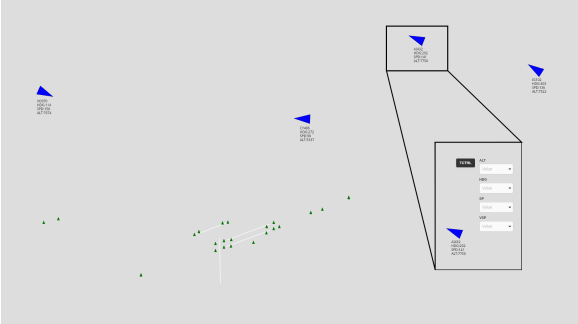

Figure 3: The simulator's user interface. The main view shows the tactical display with aircraft representations. The magnified inset highlights the graphical control panel used for issuing commands (altitude, heading, speed), demonstrating the GUI-based interaction model designed for the study.

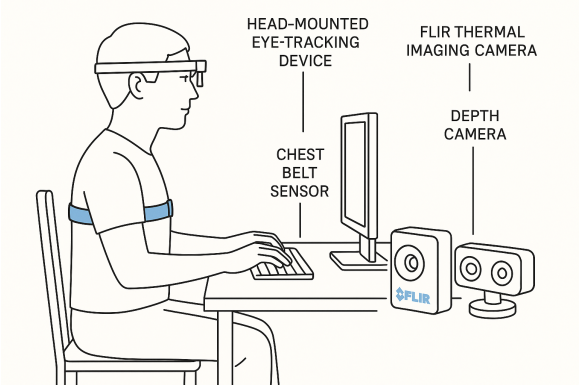

Figure 4: The experimental setup for multimodal data collection. Participants interacted with the simulator while being monitored by a head-mounted eye-tracker, a chest belt respiration sensor, a depth camera, and a FLIR thermal imaging camera to capture physiological and behavioral responses to cognitive load.

needs, such as integrating custom conflict detection algorithms. This open, modular, and extensible architecture makes BlueSky an ideal foundation for our work, providing the core simulation engine upon which we built our custom frontend and sensor integration toolkit.

The system architecture comprises two primary components: the simulation backend and the user interface frontend. The backend leverages BlueSky's network client functionality to connect with the simulation engine, exposing two WebSocket interfaces: one dedicated to streaming aircraft data and another handling control commands issued to aircraft. This design provides flexibility, allowing the backend to be hosted locally or remotely on a dedicated server.

The frontend was developed using Rust and React within a Tauri container, offering robust cross-platform compatibility. This component connects to the backend via WebSocket to receive real-time aircraft updates and send user-generated control commands. The interface's design intentionally mirrors the minimalistic aesthetics found in real-world ATC systems, rendering aircraft as blue triangles, navigational points as green triangles, and including schematic representations of airports and air traffic control sector boundaries.

Users interact with the system by taking control of aircraft entering their sector. They can then manage aircraft trajectories by adjusting altitude, speed, heading, and vertical speed through a dedicated graphical interface. The primary simulated task implemented for demonstration purposes involves managing and sequencing aircraft for landing at Frankfurt Airport.

Figures included depict the system architecture and representative screenshots of the user interface, illustrating the clarity and simplicity integral to the design.

5 Evaluation

Our evaluation process involved preliminary tests with three participants, designed primarily to validate the functionality and effectiveness of our developed toolkit. The assessment incorporated several physiological measurement instruments: a PupilLabs EyeTracker for gaze analysis, a Polar OH1 photoplethysmography (PPG) sensor for cardiovascular monitoring, a Vernier GoDirect Respiration Belt to measure breathing patterns, an Intel RealSense depth camera, and a Topdon TC001 far-infrared (FLIR) thermal camera.

The experimental protocol comprised two distinct simulation scenarios, each designed around the task of queuing aircraft for landing at Frankfurt Airport. This specific task was chosen for the initial validation as it represents a well-defined, high-workload activity that is fundamental to ATC operations, making it an ideal testbed for assessing the toolkit's ability to measure fluctuations in cognitive effort. One scenario was intentionally simplified, while the other was more complex, requiring heightened cognitive effort. Participants experienced these scenarios in a randomized sequence to mitigate order effects.

Throughout each session, participants' interactions with the simulator were meticulously recorded. Eye-tracking data were mapped directly to the screen elements to analyze attention and interaction patterns. Concurrently, physiological responses were captured using the PPG sensor, respiration belt, depth camera, and infrared camera. Specifically, facial temperature variations—measured through the combined use of the RGB data from the depth camera and thermal data from the infrared camera—enabled the detailed examination of stress-related physiological responses.

Subjective assessments complemented physiological measurements. Cognitive load was evaluated post-hoc using the NASA Task Load Index (NASA TLX) and dynamically during the simulation via periodic prompts using a 10-button interface, where participants rated their real-time cognitive load in response to audio cues.

Preliminary findings from initial test series reveal measurable physiological responses corresponding to the varying cognitive demands of the simulation tasks. Notably, respiration data from the breathing belt have emerged as particularly promising indicators of cognitive load, aligning with similar outcomes observed in earlier studies involving emergency control room simulations [15].

Figure 5: An overview of the hardware suite used for multimodal data collection. The components are (clockwise from top left): the thermal (Topdon TC001) and depth (Intel RealSense) camera assembly; the Pupil Labs head-mounted eye-tracker; the Polar OH1 photoplethysmography (PPG) armband sensor; and the Vernier Go Direct Respiration Belt.

6 Discussion

This study successfully developed and evaluated a highly realistic, sensor-enhanced Air Traffic Control simulation framework to advance Human-Computer Interaction research in safety-critical domains. Our preliminary findings confirm the framework's practical utility for capturing detailed behavioral and physiological data during complex ATC tasks. Participants were able to effectively manage aircraft sequencing under varied cognitive loads, and notably, physiological metrics such as respiration patterns showed promising correlations with task difficulty. This observation aligns with previous findings from studies in emergency control room simulations, reinforcing the potential of using multimodal sensing—including eye-tracking, photoplethysmography, and thermal imaging—to evaluate cognitive states in dynamic environments.

The adoption of the structured Design Science Research methodology was crucial for navigating the complexities of this research domain. This approach allowed us to effectively balance rigorous research goals with practical implementation constraints. To ensure the relevance of our simulation, we maintained close collaboration with an active ATC expert, which guided the design to accurately mirror real-world workflows. This collaboration also helped us make necessary research-focused adaptations, such as implementing a Graphical User Interface instead of traditional voice commands to improve data traceability. The rigor of our artifact was strengthened by integrating quantitative validation methods, such as sensor data and the NASA Task Load Index, with the qualitative feedback gathered from our expert consultant. Furthermore, the system was designed for scalability; its modular architecture, featuring the BlueSky backend and a Rust/React frontend, readily supports future extensions like incorporating new sensors or adapting the framework for other control room contexts.

Despite these strengths, our study has several key limitations. The physiological data, while consistent, is preliminary and requires validation with a larger participant pool to generalize the findings confidently. Moreover, the fusion of data from multiple sensors, such as combining thermal and RGB data for stress analysis, can introduce noise and complicate real-time interpretation. Finally, while our framework successfully captures data from multiple sensors, a significant challenge lies in the synchronization, fusion, and interpretation of these disparate data streams. The current analysis treats each sensor modality largely in isolation. The non-trivial task of time-aligning the data and developing a validated fusion model was beyond the scope of this initial work, and this represents a key barrier to achieving a single, reliable, real-time indicator of cognitive load. Furthermore, the ecological validity of the study is constrained by the use of a GUI instead of standard voice-based communication. This trade-off, made to ensure data traceability, impacts the direct applicability of our findings. The GUI may decrease certain cognitive loads (e.g., memory recall for standard phraseology) while potentially increasing others (e.g., visual search and manual dexterity tasks). Therefore, the cognitive load profiles measured in our study may not be directly comparable to those in a live ATC environment. The scope of the evaluation was also limited to approach and landing sequences at a single airport. The framework's robustness and the validity of the collected physiological metrics remain untested in other critical ATC scenarios, such as managing en-route sector handoffs, severe weather events, or unforeseen emergencies. Consequently, the current findings should be considered specific to the tested task. A primary limitation of this study is the preliminary nature of its evaluation. The findings are based on an initial test series with a small participant pool, which, while sufficient to demonstrate the framework's technical functionality, lacks the statistical power required for generalizable conclusions. The observed physiological trends, though promising, must be interpreted as illustrative rather than definitive pending validation with a larger and more diverse sample of participants, including professional controllers.

Building on this foundation, future work will focus on several key areas. To address the challenge of data integration, future work will focus on advanced analytics. We plan to develop a real-time cognitive load prediction model by applying supervised machine learning algorithms (e.g., Recurrent Neural Networks or Gradient Boosting models) to the synchronized sensor data. This process will involve critical initial steps such as feature engineering from raw physiological signals, developing robust methods for time-series alignment across sensors with different sampling rates, and validating the resulting model against established subjective measures like the NASA TLX. The framework's adaptable design also allows for domain adaptation, making it possible to apply it to other safety-critical settings, such as surgical teams or nuclear power plant control rooms. To bridge the gap between our experimental setup and operational contexts, a primary future objective is the integration of a voice-command system capable of parsing and logging commands, thereby enhancing the simulation's realism and ecological validity.

By making our simulation framework and tools openly available, we invite the research community to collaboratively advance HCI studies in safety-critical fields. Our work exemplifies how the DSR methodology can guide the development of adaptable and rigorously evaluated tools for high-stakes environments. While challenges in data interpretation persist, the integration of empirical validation with pragmatic design offers a robust pathway toward innovations that enhance both human performance and system safety.

7 Conclusion

This paper addressed the significant challenge of conducting controlled HCI studies in safety-critical environments like Air Traffic Control. We presented the design, implementation, and evaluation of a novel, low-cost simulation framework that integrates multiple physiological sensors to enable detailed cognitive load analysis. Developed using the Design Science Research methodology, our work provides a practical and extensible tool for researchers aiming to investigate operator performance in high-stakes settings.

Our preliminary evaluation demonstrated the framework's capability to capture rich, multimodal data during realistic ATC tasks. The findings confirm that physiological metrics, particularly respiration patterns, show measurable correlations with operator cognitive load, aligning with results from similar studies. This validates our approach as a viable method for objectively assessing operator states in dynamic, demanding environments.

While our findings are promising, we acknowledge the study's limitations, including its preliminary nature and the reduced ecological validity due to the GUI-based interaction. Future work will prioritize validation with larger, more diverse participant groups and enhancing realism through the integration of a voice-command system. By releasing this framework as an open-source tool, we aim to empower the research community to build upon our work, fostering collaborative efforts to develop innovations that enhance both human performance and system safety in critical domains.

References

[1] Carleton University, Qi Deng, Shaobo Ji, and Carleton University. 2018. A Review of Design Science Research in Information Systems: Concept, Process, Outcome, and Evaluation. *Pacific Asia Journal of the Association for Information Systems* (2018), 1–36. doi:10.17705/1pais.10101

[2] Guilherme da Silveira and Éder Henriqson. 2021. Design Science Research – Alternative Pathway for Aviation Training-Related Studies. *Journal of Aviation/Aerospace Education & Research* 30, 2 (Jan. 2021). doi:10.15394/jaaer.2021.1902

[3] Alan Hevner and Samir Chatterjee. 2010. *Design Research in Information Systems: Theory and Practice*. Integrated Series in Information Systems, Vol. 22. Springer US, Boston, MA. doi:10.1007/978-1-4419-5653-8

[4] Alan Hevner and Samir Chatterjee. 2010. Design Science Research in Information Systems. In *Design Research in Information Systems*. Vol. 22. Springer US, Boston, MA, 9–22. doi:10.1007/978-1-4419-5653-8_2

[5] Alan R. Hevner and Nilmini Wickramasinghe. 2018. Design Science Research Opportunities in Health Care. In *Theories to Inform Superior Health Informatics Research and Practice*, Nilmini Wickramasinghe and Jonathan L. Schaffer (Eds.). Springer International Publishing, Cham, 3–18. doi:10.1007/978-3-319-72287-0_1

[6] Johann E.W. Holm and Gabriel P.R. vd Merwe. 2019. Systems Engineering and Design Science Research for Effective Product Development. In *2019 IEEE AFRICON*. IEEE, New York City, 1–8. doi:10.1109/AFRICON46755.2019.9133820 ISSN: 2153-0033.

[7] Janine Koepp, Miriam Viviane Baron, Liane Mahlmann Kipper, Marcelo Carneiro, Mari Ângela Gaedke, Lia Gonçalves Possuelo, and Rejane Frozza. 2024. Development of an application in the healthcare area using the Design Science Research methodology. *Caderno Pedagógico* 21, 1 (jan 16 2024), 1261–1280. doi:10.54033/cadpedv21n1-066

[8] Norbert Koppenhagen, Oliver Gaß, and Benjamin Müller. 2012. Design Science Research in Action - Anatomy of Success Critical Activities for Rigor and Relevance. Poster präsentiert auf 20th European Conference on Information Systems (ECIS 2012), Barcelona, Spain, June 10-13, 2012.

[9] Gondy Leroy, Bengisu Tulu, and Xiao Liu. 2023. Introduction to the Special Issue on Design and Data Science Research in Healthcare. *ACM Transactions on Management Information Systems* 14, 2 (mar 13 2023), 1–4. doi:10.1145/3579646

[10] Ahmad M., Badr K., Colomb R., and Ibrahim R. 2012. Ontology-Based Applications In Information Systems Research: Through The Lens Of Design Science Research Methodology. *Pacific Asia Conference on Information Systems* (2012).

[11] Robert Parise, Jean-Paul Imbert, François Marais, and Roland Alonso. 2012. LABY: Un support d'aide à l'évaluation de choix de conception d'IHM pour le contrôle aérien. In *Proceedings of the 2012 Conference on Ergonomie et Interaction homme-machine*. ACM, Biarritz France, 185–188. doi:10.1145/2652574.2653433

[12] Ken Peffers, Tuure Tuunanen, Charles E. Gengler, Matti Rossi, Wendy Hui, Ville Virtanen, and Johanna Bragge. 2020. Design Science Research Process: A Model for Producing and Presenting Information Systems Research. doi:10.48550/arXiv.2006.02763 arXiv:2006.02763.

[13] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems* 24, 3 (12 2007), 45–77. doi:10.2753/mis0742-1222240302

[14] Thomas Prevot, Nancy M. Smith, Everett Palmer, Todd J. Callantine, Paul U. Lee, Joey Mercer, Lynne Martin, Connie Brasil, and Christopher Cabrall. 2014. An Overview of Current Capabilities and Research Activities in the Airspace Operations Laboratory at NASA Ames Research Center. In *14th AIAA Aviation Technology, Integration, and Operations Conference*. American Institute of Aeronautics and Astronautics, Atlanta, GA. doi:10.2514/6.2014-2860

[15] Jonas Pöhler, Antonia Vitt, Nadine Flegel, Tilo Mentler, and Kristof van Laerhoven. 2023. Investigating Cognitive Load in Emergency Control Room Simulations. (2023). doi:10.18420/MUC2023-MCI-WS01-355

[16] Jeff Scales. 2020. A design science research approach to closing the gap between the research and practice of project scheduling. *Systems Research and Behavioral Science* 37, 5 (9 2020), 804–812. doi:10.1002/sres.2743

[17] Vijay K. Vaishnavi. 2007. *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*. Auerbach Publications, New York. doi:10.1201/9781420059335

[18] John Venable and R. Baskerville. 2012. Eating our own cooking: Toward a more rigorous design science of research methods. *Electronic Journal of Business Research Methods* 10 (01 2012), 141–153.