Context-Aware Human Performance Measurement for Simulation-based Tactical Training

Joost van Oijen, Thomas Bellucci, Maxim van Oldenbeek
NLR – Royal Netherlands Aerospace Centre
Amsterdam, the Netherlands
Joost.van.Oijen@nlr.nl, Thomas.Bellucci@nlr.nl,
Maxim.van.Oldenbeek@nlr.nl

ABSTRACT

Modern military forces increasingly introduce simulation-based training to develop and sustain critical skills through procedure, tactical, or mission training in realistic threat environments. However, effectively assessing trainee performance remains a challenge. Most simulation systems include data recording, monitoring, or debrief/after-action review (AAR) tools to support trainee performance measurement and assessment. While these tools capture relevant data required for performance measurement, they often do not utilize this data for objective, actionable insights for human assessors, such as instructors. The usage of captured data is usually limited to basic visualizations or communication replays, leaving contextual interpretation and analysis concerning trainee task performance to the human assessor.

In this paper, we present a lightweight measurement framework for real-time context-based performance analytics to support trainee assessment in simulation-based training. The framework supports three key features. First, it is based on the chaining of individual measures to create a network of progressively rich information, ranging from (1) ground-truth data to (2) task-specific performance metrics and (3) objective assessments. Second, it supports the contextual activation of measures such that measurements can be delivered in real-time at the right time, relevant to the current task, situation, or mission phase. Finally, we show how AI techniques, such as Deep Learning for behavioral recognition, Natural Language Processing for communication analysis, and rule-based methods for procedural evaluation, can enhance performance assessment. The outcome of measures can be visualized and delivered to human assessors by means of a dashboard, or fed into instructional systems for learner profile management.

We evaluate the framework in a case study for fighter pilot training. Based on a use case scenario, we demonstrate how our framework enhances instructor capabilities by providing real-time, data-driven insights into pilot performance across key skill areas, including communication, tactical decision-making, and maneuvering.

ABOUT THE AUTHORS

Joost van Oijen, PhD. is a Senior Scientist at the Royal Netherlands Aerospace Centre (NLR) at the department of Training & Simulation. He obtained his MSc. in Computer Science in 2007 and his Ph.D. in Artificial Intelligence in 2014. At his current position, Joost leads several Defense research projects focused on Augmented Intelligence and human behavior modeling across domains of training, decision support and human-machine teaming.

Thomas Bellucci, MSc. is an R&D engineer at the Royal Netherlands Aerospace Centre (NLR) at the department of Training & Simulation and works on research projects involving Artificial Intelligence (AI), data science and machine learning. He studied Artificial Intelligence at the University of Amsterdam (UvA) and the Vrije Universiteit (VU), Amsterdam, completing both his BSc and MSc programs with distinction (summa cum laude).

Maxim van Oldenbeek, MSc. is an R&D engineer at the Royal Netherlands Aerospace Centre (NLR) at the department of Training & Simulation. He holds two master degrees: one in Mathematical Sciences from Utrecht University and another in Data Science from the University of Copenhagen. In his current position Maxim contributes to a variety of AI-driven projects.

Context-Aware Human Performance Measurement for Simulation-based Tactical Training

Joost van Oijen, Thomas Bellucci, Maxim van Oldenbeek
NLR – Royal Netherlands Aerospace Centre
Amsterdam, the Netherlands
Joost.van.Oijen@nlr.nl, Thomas.Bellucci@nlr.nl,
Maxim.van.Oldenbeek@nlr.nl

INTRODUCTION

Simulation-based training is increasingly used by defense organizations to train military personnel in a safe and cost-effective manner. In the aviation domain, for example, fighter and helicopter pilots undergo simulation-based training to develop and maintain job-specific competencies through tactical training in simulated threat environments. Furthermore, there is a growing interest in mission training to increase warfighter readiness in dynamic, operationally relevant scenarios, supported by advanced simulation environments, such as Mission Training through Distributed Simulation (MTDS) and Live-Virtual-Constructive (LVC) setups (Lemmers & Petermeijer, 2022).

Many modern simulation facilities used for training are equipped with data recording facilities and offer real-time monitoring and debrief tools. These tools aid instructors and trainees in identifying performance deficiencies, training gaps, or lessons learned, by observing trainee performance during training or after action review (AAR) (Hanoun & Nahavandi, 2018). For instance, fighter pilots often rely heavily on AAR for learning, by revisiting and analyzing critical situations and decision points from training sessions (Aronsson et al., 2019). In this context, the use of computer-assisted performance analytics, supported by Data Science and Artificial Intelligence (AI), can greatly enhance in-training or AAR learning. Tailored analytics can provide objective, contextualized insights into technical or non-technical skills, such as motor skills, tactical decision-making, communication, or teamwork.

Although many off-the-shelf monitoring or debrief tools capture relevant simulation data required for performance analytics, they often fall short in utilizing this data to create meaningful insights. Typically, data usage is limited to visualizing raw telemetry or replaying recorded communications, leaving the interpretation and contextual analysis of trainee performance to human observers. General-purpose tools are rarely equipped with built-in analytics for specific training tasks, as such analyses quickly become dependent on governed nation- or organization-specific knowledge, such as doctrine, Tactics, Techniques, and Procedures (TTPs), or tactical operating procedures. Furthermore, such knowledge may be treated as sensitive or classified, hindering the development of general-purpose algorithms.

In this paper, we present a lightweight framework for computer-assisted performance measurement in simulation-based tactical training. The purpose of the framework is to enable efficient development and deployment of domain-specific performance analytics that (1) interfaces easily with existing simulation environments via standard data protocols, (2) can provide real-time, context-relevant insights to end users, and (3) can be agnostic to underlying required AI technologies. The framework's capabilities are demonstrated through a case study in fighter pilot training, focusing on the assessment of various technical skills. This paper focuses on the technical aspects of the framework, where investigated analytics have been demonstrated in an agent-based simulation environment, representative of a human-in-the-loop training environment. The training value of investigated analytics has not been validated in human experiments, which is left for future work. Looking ahead, the framework will serve as a foundation for future research directions, where the aim is to explore the use of generative AI for the automated generation and real-time injection of analytics based on user demand.

The remainder of this paper is structured as follows: Section 2 describes the conceptual foundation of computer-assisted performance measurement. Section 3 introduces the case study in fighter pilot training with three implemented use cases. Section 4 presents the underlying implementation of the framework. Section 5 concludes with a summary of findings and directions for future work.

COMPUTER-ASSISTED PERFORMANCE MEASUREMENT

Simulation-based training enables warfighters to maintain combat readiness through scenario-driven exercises. In tactical training, participants learn to apply skills and procedures in mission-oriented tactical situations that emphasize situational awareness, decision-making, communication, and teamwork. Examples include a team of soldiers practicing clearing a building, or fighter pilots rehearsing air combat game plans. To identify readiness gaps and performance deficiencies, computer-assisted performance measurement can support trainee assessment by providing objective, data-driven insights. These insights can support instructors or trainees during training or debrief by complementing subjective human observations with quantitative analysis (van Oijen, 2024). Performance data can also be integrated into instructional systems or learning ecosystems to manage learner profiles and inform instructional planning (B. Smith & Milham, 2021).

Computer-assisted performance measurement relies on learning analytics, defined as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Long, 2011). In the context of simulation-based military training, (Watz et al., 2023) discusses the growing interest in tools and techniques for applying learning analytics to transform performance data into actionable insights. The authors identify three key enablers: (1) suitable *data infrastructures* for collecting performance data, processing analytics, and delivering results to end users, (2) *domain knowledge engineering* to develop objective analytics for meaningful training intelligence, and (3) *contextual understanding* to interpret performance data for meaningful assessment. In this study, we address these three pillars as core features in the development of a lightweight computational framework to support real-time performance measurement in simulation-based tactical training. A high-level overview of this framework is presented in Figure 1.

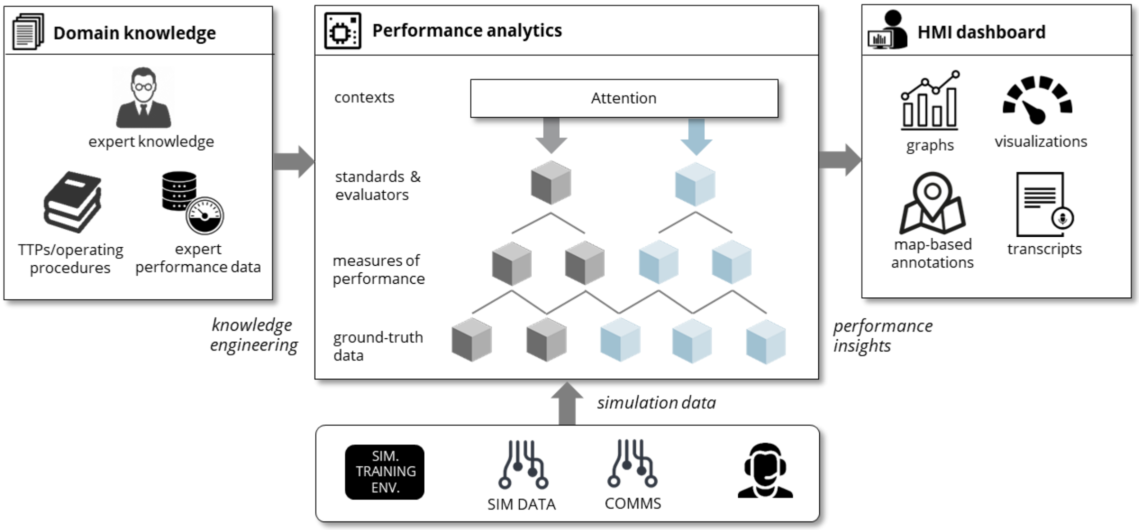


Figure 1: Computer-assisted performance measurement: the application of performance analytics (center) to deliver objective performance insights about trainees in a simulation-based training environment (bottom) towards human assessors through human machine interfaces (HMI) (right), by exploiting objective domain (expert) knowledge (left).

The framework serves to automate aspects of performance assessment through computational algorithms, providing end users with (real-time) objective data on trainee performance. To deliver effective support within a specific training domain, e.g., fighter pilot training, requires addressing three foundational questions, discussed in the remainder:

- What kind of performance analytics can be objectively determined?
- When to apply performance analytics such that they are contextually relevant?
- How to embed performance analytics in simulation-based training environments?

Objective performance analytics

Automated performance assessment relies on data from the training environment to evaluate trainees' task execution against an objective standard for that task. This requires the ability to encode and represent an objective performance standard into a computational model that can act as a critic. The feasibility of such an assessment depends on whether such a standard exists, whether expert knowledge defining it can be acquired, and whether it can be formalized algorithmically. As illustrated in Figure 1, domain knowledge may originate from explicit documents, (implicit) human expert knowledge, or example performance data (e.g., recorded from live missions or simulated exercises).

An appropriate algorithmic approach for automated assessment depends on whether the task involves *explicit* or *tacit* knowledge (E. A. Smith, 2001). For tasks grounded in explicit knowledge, criteria for performance can be articulated, communicated, and captured in decision rules. Examples include specific tactics to follow, adherence to combat parameters, or using communication protocols. In defense organizations, such knowledge may be secured in documents such as TTPs or Tactical Operating Procedures (TOPs) as unit-specific rules for executing combat tasks. Furthermore, explicit performance criteria may be communicated before training during the briefing. However, documented or communicated criteria may only apply to a limited set of scenarios, leaving gaps when trainees face unforeseen situations where no predefined correct performance exists.

In contrast, for tacit knowledge tasks, performance criteria are harder to express or codify in tangible form, relying more on the experience or intuition of experts. Examples include tasks that require motor skills such as dogfight maneuvering or helicopter landing, or non-technical tasks like problem-solving or teamwork. Two main approaches support their assessment, namely knowledge-based and data-driven approaches. Knowledge-based methods may identify quantifiable Measures of Performance (MoPs) through task analysis or expert input. For example, MoPs have been developed for helicopter maneuvers (Thijssen et al., 2024) and team-based room-clearing tasks (Vatral et al., 2022). Alternatively, data-driven approaches use examples of behavior traces to train models using machine learning. For example, Stevens-Adams et al. (2010) trained a system to assess tactical air engagements based on human-graded examples, showing strong alignment with expert ratings. Bewley et al. (2024) applied reinforcement learning from human feedback (RLHF) to train expert models of fighter pilot behaviors, enabling the encoding of tacit expert preferences used for automated assessment or demonstration learning.

Context-aware performance analytics

Understanding the situational context of learners is essential for accurate performance assessment, as emphasized in the earlier definition of learning analytics. Contextual understanding involves two elements: (1) the scope of information required to evaluate a behavioral task, and (2) the conditions under which specific behaviors should be measured. Here, we focus on the latter. In military tactical training, for example, decisions and actions are best evaluated in relation to mission phases, tactical situations, or the specific tasks being performed. Accurately recognizing such contexts is critical for timely and meaningful interpretation of performance data. While human observers naturally apply contextual judgment, computational systems must encode this understanding explicitly. To address this need, the framework includes a top-down attention model (see Figure 1), which guides the activation of relevant performance measures based on situational relevance. This model ensures that metrics are computed only when contextually appropriate (e.g., assessing a tactical maneuver only during its execution), aligned with current tactical situations, and meaningful to the end user.

Environment embedding

To embed (objective and context-aware) performance analytics within a training environment, a suitable data infrastructure is needed. In earlier work (van Oijen, 2024), we distinguished three data processing levels to support end users—depicted in Figure 1 as an interconnected network of measures that build progressively richer information. The first level collects raw (multimodal) data as the ground truth from the simulation environment. Common data sources in military simulation are provided by protocols such as Distributed Interactive Simulation (DIS) (IEEE, 2012), delivering real-time information about entities and events according to a standardized semantic object model. Some data modalities require further processing—for example, using automatic speech recognition (ASR) to infer transcribed text from embedded voice data. The second level transforms ground-truth data into task-specific MoPs, yielding quantifiable performance indicators. The third level consists of evaluators who compare measured performance data to standards to provide objective assessments. While higher levels offer more direct value to end users, they also require greater engineering complexity, domain expertise, and contextual awareness (Choudhary, 2024). As fully automated assessment is not always feasible or practical, insights at lower levels (e.g., ground truth data or MoPs) remain valuable as objective data for end users to create subjective assessments. The data infrastructure should therefore support the delivery of performance data at any level. Figure 1 illustrates various HMI components that present performance data to end users, including map-based annotations, voice transcripts, and graphs.

CASE STUDY

This study presents a lightweight implementation of the conceptual performance measurement framework described in the previous section. A case study was used to evaluate the framework based on an implemented proof-of-concept application in the domain of fighter pilot training. It was designed to cover relevant requirements through different use case examples of performance assessment. These address the need for assessing both explicit and tacit knowledge tasks, assessment based on multi-modal data, and context-aware measurement. Furthermore, its implementation requires domain knowledge engineering from various military knowledge sources, as well as the use of various AI approaches. In the remainder of this section, the case study is described in detail. The technical implementation of the framework to support the case study implementation is then presented in the next section.

The case study focuses on a Beyond Visual Range (BVR) air combat training scenario[1]. In the early stages of training, pilots can practice standard tactical scenarios to apply procedures and skills in line with their organization's TTPs. These *standard* scenarios cover common threat situations for which well-defined performance criteria are documented, making them suitable for automated assessment. This contrasts with *non-standard* or unforeseen situations, which lack precise assessment criteria and require trainees to rely more on experience, intuition, or creativity. A monitoring or debrief tool capable of providing objective performance insights in standard scenarios enhances opportunities for self-directed learning, reducing reliance on instructor interpretation.

Scenario

The training scenario resembles an air combat engagement in a standard 2v2 encounter and is structured around three key phases, illustrated in Figure 2. Each phase corresponds to a distinct use case that targets a specific skill domain—communication, tactical decision-making, and maneuvering—performed by different roles: the Fighter Controller (FC), Flight Lead (FL), and Wingman (WM), respectively.

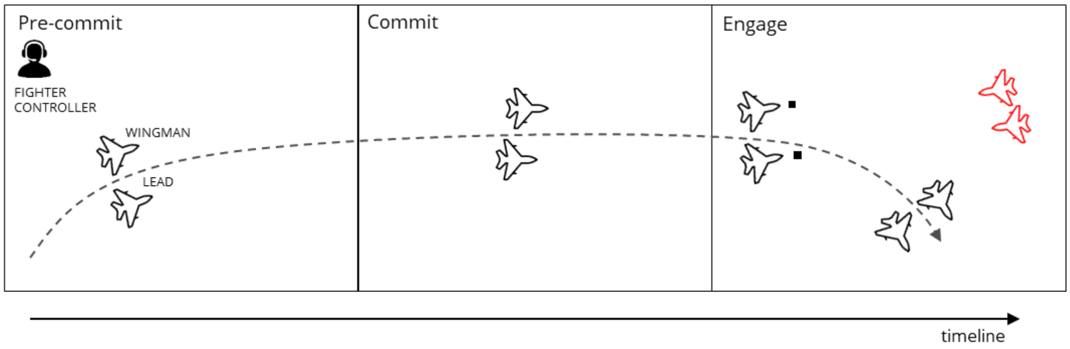


Figure 2: Phases in an air combat engagement

The *pre-commit* phase involves the initial detection of enemy aircraft. Here, the FC is responsible for recognizing and communicating threat information, a so-called red air picture. This use case exemplifies situational awareness and communication skills. The *commit phase* focuses on the FL, who must implement a targeting plan to assign specific threats to each team member—a tactical decision-making task. Finally, in the *engage phase*, both the FL and WM perform one-on-one maneuvers against respective threats. This use case targets the measurement of combat maneuvering proficiency.

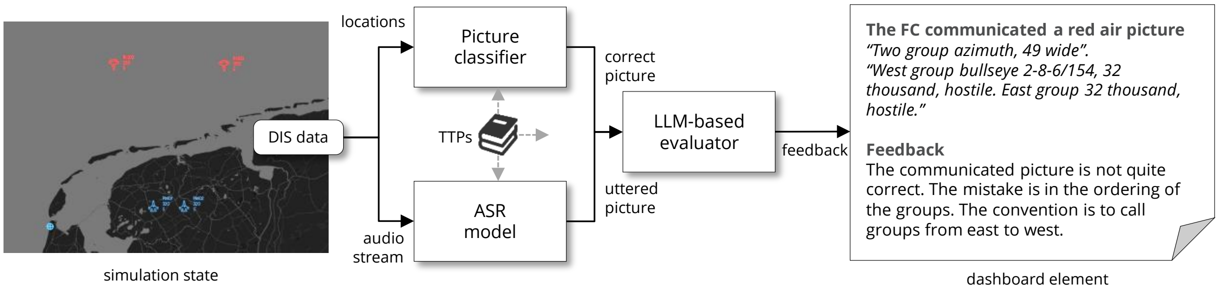
To support the measurement of these tasks throughout a training session, context-aware measurement is needed. In air combat, engagement phases can be inferred from spatial relationships, specifically, the distance to enemy groups. As aircraft cross predetermined (and often classified) distance thresholds, specific tasks become active for each role. This knowledge is used to define three engagement contexts and their transitions, in line with the described phases. During training, only measurements relevant to the current context are processed and presented to the end user. The following subsections detail the implementation of each use case, describing the assessed tasks, the origin of the performance standard, the automated measurement approach, and how results are presented to the end user.

[1] Any military information referenced in this section is illustrative and has been simplified or modified to ensure suitability for public dissemination. No sensitive or classified content is included.

Use case 1: Picture Communication

Task and performance standard: This use case illustrates a situational awareness and communication task, specifically, the fighter controller's assessment and communication of a so-called 'red air picture' (*pre-commit* phase from Figure 2). A red air picture is a cognitive representation of enemy threat locations communicated verbally between fighter controllers and pilots. It conveys information such as enemy formations and sizes, group labeling, and locations relative to briefed airspace. The communication of a picture is an explicit knowledge task, with performance criteria's well-defined in an organization's TTPs on air communication standards.

Computer-assisted assessment: Since the task involves voice communication, its performance is directly observable from the environment. For the assessment, we compare the spoken description of a red air picture uttered by the trainee against the correct description according to the communication standard (see Figure 3). To represent the performance standard, we constructed a rule-based model, referred to as the Picture Classifier, by formalizing the semantics and rules from a doctrinal document on communication standards (i.e., TTPs) into a computational model. The model composes a picture definition based on observed positions of blue and red air from positional data and determines the corresponding surface-level text following communication standards. To measure the trainee's behavior, we employed an automatic speech recognition (ASR) model that transcribes real-time speech to text using OpenAI's open-source Whisper model. Given that picture communication uses specialized jargon rarely present in general speech datasets used for training ASR models, we applied prompting techniques in Whisper to improve recognition accuracy and reduce errors. Finally, a large language model (LLM) provides feedback and explanations for discrepancies between the trainee's communication and the standard. Currently, the LLM implementation relies on predefined explanation templates as a proof-of-concept. Future enhancements could include embedding doctrinal documents into the LLM via prompting or fine-tuning approaches to enable more intelligent, domain-specific feedback and explanation.



The FC communicated a red air picture
"Two group azimuth, 49 wide". "West group bullseye 2-8-6/154, 32 thousand, hostile. East group 32 thousand, hostile."

Feedback
The communicated picture is not quite correct. The mistake is in the ordering of the groups. The convention is to call groups from east to west.

Figure 3: Assessment method for the red air picture communication task. Detected picture communication (ASR) is compared against a standard (the classifier), to deliver an assessment of performance (the evaluator). TTP documents are employed for knowledge engineering across all measures, to offer a robust standard, ensure accuracy of speech recognition, and generate explanatory feedback.

Measurement support: To assist the end user, developed measures are presented via dashboard elements. The Picture Classifier's output can display the performance standard either as text or through annotations on a mini-map. The evaluator's output can provide LLM-generated assessment feedback using text. Additionally, the ASR output provides transcripts of communication occurring throughout the training session.

Use case 2: Targeting

Task and performance standard: The second use case focuses on assessing the flight lead's tactical decision-making during the commit phase, specifically the task of targeting. In this task, the flight lead must decide which team member should engage which enemy aircraft (*commit* phase in Figure 2). This represents another task grounded in explicit knowledge, with performance standards defined by a decision-making framework derived from TTPs. These procedures specify targeting rules based on factors such as relative positioning, proximity, and capabilities of individual threats, providing an objective basis for assessment.

Computer-assisted assessment: To assess the flight lead's targeting plan, we analyze audio communications between the flight lead and the wingman to extract the implied assignment of blue forces to red targets. This inferred targeting plan is then compared to an ideal plan derived from the TTP-based decision model, which calculates optimal assignments using scenario-specific data (e.g., positions, distances, and identities). This mirrors the approach in use case 1, namely using ASR for measuring communication, a rule-based algorithm to calculate the ideal plan for the given situation (the standard), and an LLM-based evaluation to compare the communicated plan to the standard.

Measurement support: To support the end user, the resulting performance measures are presented through dashboard visualizations. These highlight how closely the flight lead's plan aligns with doctrinal procedures by marking any discrepancies between the communicated targeting plan and the one prescribed by the TTPs.

Use case 3: Maneuvering

Task and performance standard: This final use case involves assessing a maneuvering task as a technical skill. Specifically, it concerns an offensive maneuver during an enemy engagement (*engage* phase from Figure 2), performed after launching a semi-active radar-homing missile. The firing aircraft needs to illuminate the target aircraft with its radar for the missile to be guided towards its target, while at the same time keeping a maximum distance from it. This is a balancing act where the aircraft seeks the optimal turn angle while keeping the enemy on radar, often in response to enemy counter-maneuvers. This use case illustrates a tacit knowledge task, which can be trained through repeated experience. Although explicit performance metrics are available in TTPs, such as the optimal turn angle, an overall performance standard for maneuvering proficiency cannot be articulated and exists only as implicit knowledge in the minds of trained experts.

Computer-assisted assessment: The maneuvering performance is fully observable from the environment through movement and positioning data from both the performing and target aircraft. In order to measure the performance of this tacit knowledge task, we developed a supervised machine learning model as a performance benchmark that is able to assess and score an executed maneuver. The model is trained using a dataset of example maneuvers. As human expert examples are scarce (e.g., from real-world mission data) or resource-intensive to obtain (e.g., from expert demonstrations in simulation), synthetic data generation was employed to develop the dataset. The approach is shown in Figure 4. Synthetic data is generated from examples of agent-based execution of the maneuver in a (fast-time) Computer Generated Forces (CGF) simulation. Tailored scenario generation is used to develop a balanced dataset of tactical situations in which the maneuver is performed[2]. An evaluation of the model performance falls outside the scope of this paper.

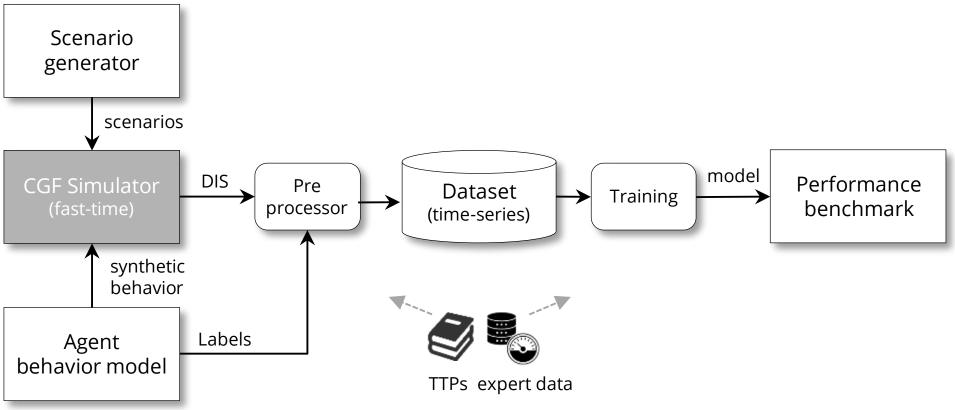


Figure 4: Method for generating synthetic data to train a performance model as a benchmark for assessment. An agent behavior model informed by TTP is used to generate synthetic behavior traces of a particular task to create a dataset of representative expert performance data, consequently used to train a performance model.

[2] As a proof-of-concept, the agent-based maneuver execution is driven by a rule-based behavior model, employing performance parameters documented in TTPs. Alternatively, the behavior model could be learned using reinforcement learning where performance parameters are incorporated in the reward function.

Measurement support: To support the end user, developed measures are presented via dashboard elements. The ML-based performance benchmark model provides a proficiency score, while a complementary rule-based measure calculates deviation from optimal turn angles defined in TTPs. These outputs are visualized through scoring indicators and graph widgets. Additionally, the optimal maneuver determined by the benchmark could be shown as a 'ghost' trajectory on a map to provide visual insight on expert performance.

FRAMEWORK IMPLEMENTATION

The proof-of-concept application with the three use cases from the previous section was built using an implementation of the performance measurement framework. It was implemented as a lightweight software framework developed in Python, designed to collect real-time data from a simulation environment, and supports both real-time and post-session performance analytics through user-defined measures. An illustration of the framework implementation is shown in Figure 5.

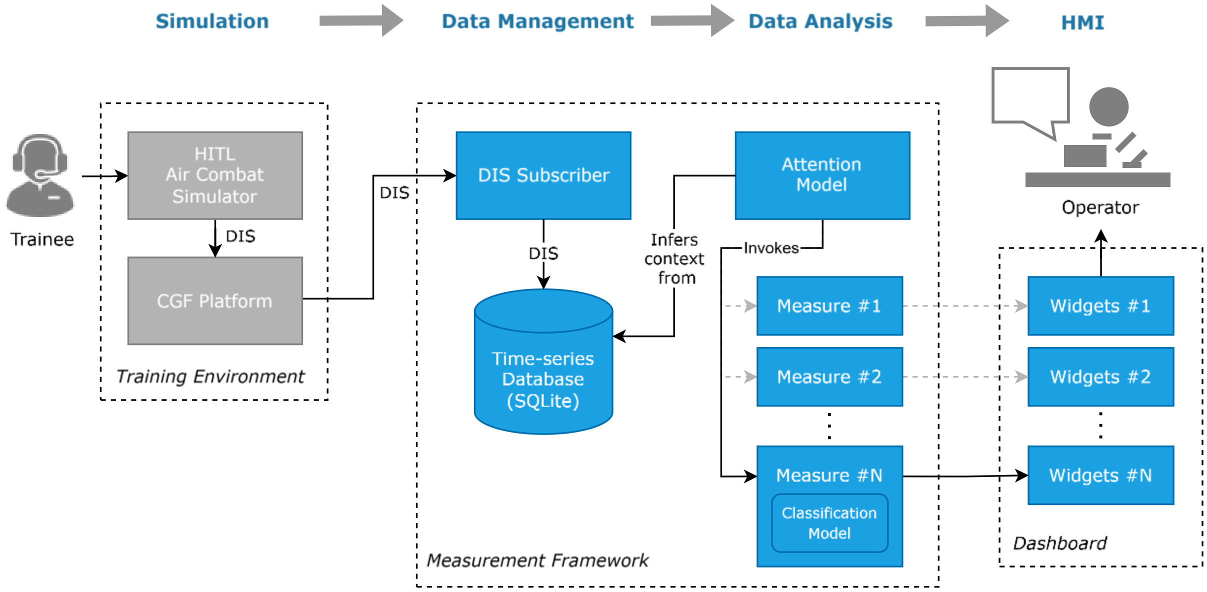


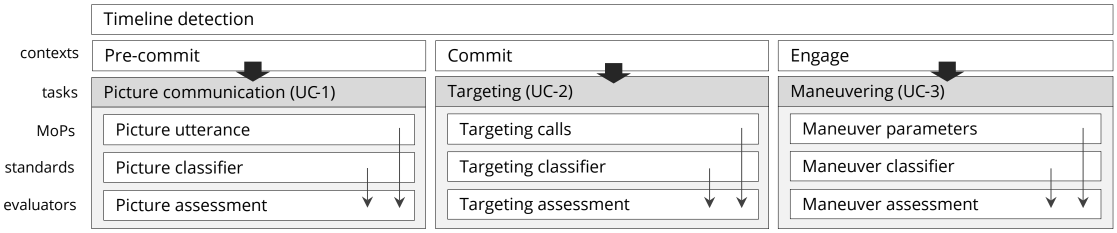
Figure 5: Framework implementation with the Training Environment providing simulation data (left), the Measurement Framework supporting performance analytics (center), and a Dashboard for end users (right)

Data management layer

The data management layer captures ground-truth information from the simulation environment over the course of a training session. For this prototype, DIS was used as the primary data source (IEEE, 2012), which was sufficient to support the skills assessments from the case study. It delivers data on e.g., aircraft identities, positioning and attitudes, launched missile events, and voice radio communication. Future versions may include additional data sources, such as tactical data link communication if supported by the simulation environment. Captured DIS Protocol Data Units (PDUs) are stored in a time-series SQLite database. During the data collection process, basic first-order features are computed and recorded, such as acceleration, Mach speed, and G-forces. The database can be accessed by downstream measures to calculate metrics of interest for performance analysis.

Measures

A *measure* is a modular analytical unit that performs a user-defined function in support of performance assessment. Measures can be chained to build upon the output of other measures, thereby allowing progressively rich information to service different levels of analytics: they can represent task-related MoPs, objective performance standards, or evaluations, as introduced in Figure 1. By extending a simple base class, user-defined measures can be developed to implement task-specific analytics and satisfy the specific information needs for a particular task domain. The implementation of measures can range from basic calculations (e.g., the distance between two aircraft) to more complex AI-based models (e.g., the deep learning maneuver classification algorithm from Use Case 3). Figure 6 shows the range of measures implemented to support the use cases. Several additional lightweight measures have been implemented to provide basic information relevant for analyzing air combat situations, such as counting missile shots, tracking distance thresholds, and identifying target aspects (e.g., hot, beam, flank, cold) during engagements.



Timeline detection

contexts	Pre-commit	Commit	Engage
tasks	Picture communication (UC-1)	Targeting (UC-2)	Maneuvering (UC-3)
MoPs	Picture utterance	Targeting calls	Maneuver parameters
standards	Picture classifier	Targeting classifier	Maneuver classifier
evaluators	Picture assessment	Targeting assessment	Maneuver assessment

Figure 6: User-developed measures for the case study. A timeline detection measure detects distance thresholds to activate various engagement contexts, which activate task-related MoPs, standards and evaluators for use cases 1-3. The narrow arrows represent dependencies between measures.

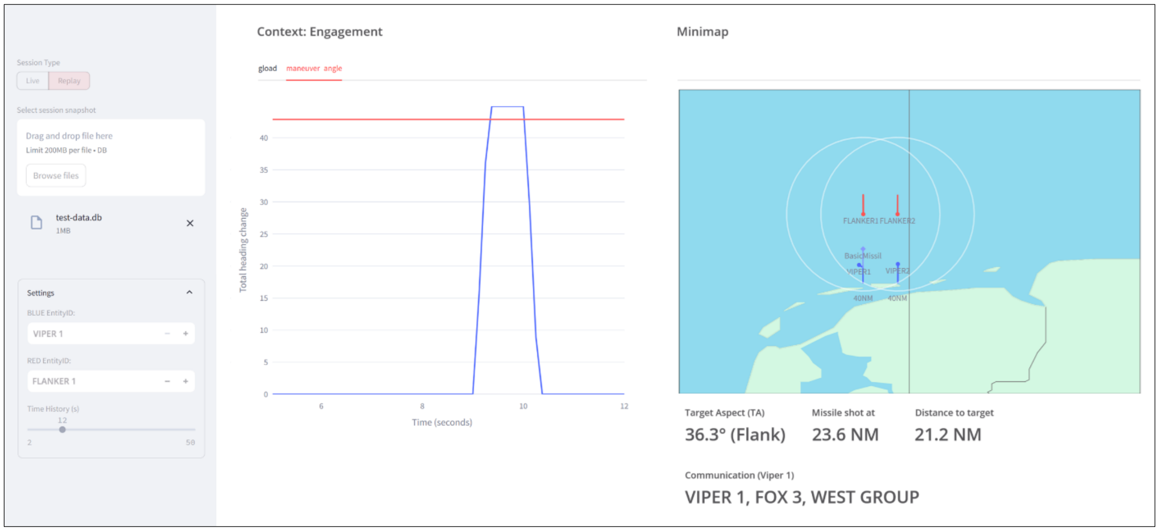
Attention model

The attention model is responsible for context-aware performance measurement by activating and deactivation measures based on current relevant situational contexts. Context inference and transitions are based on a user-defined context topology, inspired by the Context-based Reasoning (CxBR) modelling paradigm employed in an earlier study (van Oijen, 2022). In the case study, the context topology reflects the different engagement contexts shown in Figure 6. A default global context is also available for global measures that are always active. An example is the transcription measure that transforms voice data to surface-level text using ASR and speaker identification.

Multiple context topologies can run in parallel, each scoped to specific individuals or teams in different unit hierarchies (e.g., elements, flights, or squadrons). This supports the concurrent tracking of performance of multiple participants in the same team, or parallel air combats performed by different teams, operating simultaneously at potentially different geographical locations involving different enemies. In our case study, this is shown by the parallel tracking of maneuver performance of both the Flight lead and Wingman from use case 3.

Human machine interface

To deliver real-time, context-aware performance insights to the end user, a proof-of-concept dashboard was developed as the HMI. The dashboard was implemented in Python using *Plotly*, a library for building interactive data visualizations. Figure 7 shows a snapshot of the dashboard during an individual 1v1 engagement. The interface includes graphs, map annotations, and textual widgets to present the task-specific performance metrics. A control panel shown on the left allows switching between a live mode for operating on real-time session data, and a replay mode for debrief, based on stored session recordings. Furthermore, it allows selecting the trainee of interest for role-specific insights, such as for the individual actors from the case study.



Session Type: Replay

Select session snapshot

Drag and drop file here
Limit 200MB per file • DB

test-data.db
1MB

Settings

BLUE EntityID: VIPER 1

RED EntityID: FLANKER 1

Time History (s): 12 (2–50)

Context: Engagement

Target Aspect (TA)	Missile shot at	Distance to target
36.3° (Flank)	23.6 NM	21.2 NM

Communication (Viper 1)
VIPER 1, FOX 3, WEST GROUP

Figure 7: Proof-of-concept dashboard application

Enabling AI-driven performance measurement

The framework is agnostic to whether behavioral data originates from human trainees or synthetic agents (e.g., CGF). As a result, the same analytics can be applied to virtual actors and can be used for instance to facilitate testing and validation of agent behaviors against specific performance criteria. Furthermore, the framework facilitates the development of various AI-based measures. For example, captured data can be used to generate datasets that contain both human or synthetic (agent) behavior traces. As demonstrated in Use Case 3, synthetic data was used to train a deep learning classifier for maneuver recognition (see Figure 4). Because the same data model is used for training and deployment, trained models can be embedded directly into user-defined measures with minimal feature engineering. In a similar manner, reinforcement learning (RL) approaches can be embedded, for instance to represent performance standards for tactic knowledge tasks, where existing performance measures could be reused during RL training to guide the learning process.

Concluding

The implemented infrastructure offers a lightweight, developer-friendly framework that can be easily replicated and applied to existing military simulation environments to enhance training through computer-assisted performance measurement in specific task domains. It supports rapid prototyping, integration, and replacement of user-defined measures, enabling delivery of domain-specific, role-specific, and mission-level insights across individual and team actors in the simulation.

CONCLUSION

This paper presented a framework for computer-assisted performance measurement to support trainee assessment in simulation-based training. The framework’s modular structure enables the development of real-time performance analytics across different levels, supporting integration with monitoring dashboards and debrief tools.

We demonstrated the framework in a case study of tactical training in the air combat domain, using a CGF simulation platform that relied solely on the DIS protocol for acquiring real-time simulation data. Three use cases illustrated performance analytics applied to key skill areas: communication, tactical decision-making, and maneuvering proficiency. A proof-of-concept application highlighted three core capabilities introduced at the outset: (1) supporting analytics through domain knowledge engineering, (2) enabling context-aware analytics, and (3) providing a flexible infrastructure for developing and delivering analytics to end-users. Domain knowledge engineering for performance assessment included both a knowledge-driven approach (based on external documents for explicit knowledge tasks) and a data-driven approach (using synthetic expert data for tacit knowledge tasks). Context-aware analytics were demonstrated through adaptive user interfaces that tailored measurement support to mission phases and scenario activities. The framework’s flexibility was further shown through the integration of diverse AI techniques, including rule-based approaches, natural language processing, LLMs, and deep learning.

Across military training domains, organizations maintain reference materials that capture valuable domain knowledge for performance assessment. Combined with input from subject matter experts, this knowledge supports the development of performance analytics for particular training domains and tasks. In our study, this knowledge was largely manually encoded into performance constructs (e.g., MoPs, standards, assessments). While effective, this knowledge engineering process requires considerable development effort and may not always match dynamic user requirements for performance insights. To address these limitations, future work will explore the use of LLMs in two ways. First, we aim to employ them for (semi-)automated knowledge engineering. LLMs could help extract structured measures (such as MoPs, performance standards, and evaluations) from organizations’ natural language sources. Within our framework, computational measures share a unified logical structure, making it feasible to apply code generation techniques to automatically produce and deploy these measures. Second, we aim to enable user-driven analytics by allowing end-users to request insights via natural language. LLMs could translate these queries into executable measures, with results visualized using appropriate data representations. These directions align with broader trends in data analytics and offer promising opportunities to advance computer-assisted performance measurement in military training.

REFERENCES

Aronsson, S., Artman, H., Lindquist, S., Mitchell, M., Persson, T., Ramberg, R., Romero, M., & Vehn, P. ter. (2019). Supporting after action review in simulator mission training: Co-creating visualization concepts for training of fast-jet fighter pilots. *The Journal of Defense Modeling and Simulation*, *16*(3), 219–231. https://doi.org/10.1177/1548512918823296

Bewley, T., Lawry, J., & Richards, A. (2024). Learning Interpretable Models of Aircraft Handling Behaviour by Reinforcement Learning from Human Feedback. *AIAA SCITECH 2024 Forum*, 1380.

Choudhary, T. (2024). Domain Expertise in Data Analytics: Enhancing Insights across Industries. *International Journal of Research in Computer Applications and Information Technology*, *7*(2), 69–82.

Hanoun, S., & Nahavandi, S. (2018). Current and future methodologies of after action review in simulation-based training. *2018 Annual IEEE International Systems Conference (SysCon)*, 1–6.

IEEE. (2012). IEEE Standard for Distributed Interactive Simulation–Application Protocols. *IEEE Std 1278.1-2012 (Revision of IEEE Std 1278.1-1995)*, 1–747. https://doi.org/10.1109/IEEESTD.2012.6387564

Lemmers, A., & Petermeijer, B. (2022). LVC for Joint and Combined Air Power. *Proceedings of the MSG-197 Symposium on Emerging and Disruptive Modelling and Simulation Technologies to Transform Future Defence Capabilities*.

Long, P. (2011). *LAK'11: Proceedings of the 1st International Conference on Learning Analytics and Knowledge, February 27-March 1, 2011, Banff, Alberta, Canada*. ACM.

Smith, B., & Milham, L. (2021). Total Learning Architecture (TLA) Data Pillars and Their Applicability to Adaptive Instructional Systems. In C. Stephanidis, D. Harris, W.-C. Li, D. D. Schmorrow, C. M. Fidopiastis, M. Antona, Q. Gao, J. Zhou, P. Zaphiris, A. Ioannou, R. A. Sottilare, J. Schwarz, & M. Rauterberg (Eds.), *HCI International 2021—Late Breaking Papers: Cognition, Inclusion, Learning, and Culture* (pp. 90–106). Springer International Publishing.

Smith, E. A. (2001). The role of tacit and explicit knowledge in the workplace. *Journal of Knowledge Management*, *5*(4), 311–321.

Stevens-Adams, S. M., Basilico, J. D., Abbott, G., Robert, Gieseler, C. J., & Forsythe, C. (2010). Performance Assessment to Enhance Training Effectiveness. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL*.

Thijssen, D., de Marez Oyens, P., & van der Pal, J. (2024). Navigating the Skies: Enhancing Military Helicopter Pilot Training Through Learning Analytics. In R. A. Sottilare & J. Schwarz (Eds.), *Adaptive Instructional Systems* (pp. 72–88). Springer Nature Switzerland.

van Oijen, J. (2022). Human Behavior Models for Adaptive Training in Mixed Human-Agent Training Environments. *Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*.

van Oijen, J. (2024). Augmented Intelligence for Instructional Systems in Simulation-Based Training. In R. A. Sottilare & J. Schwarz (Eds.), *Adaptive Instructional Systems* (pp. 89–101). Springer Nature Switzerland.

Vatral, C., Biswas, G., Mohammed, N., & Goldberg, B. S. (2022). Automated Assessment of Team Performance Using Multimodal Bayesian Learning Analytics. *Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*.

Watz, E., Neubauer, P., Shires, R., & May, J. (2023). Precision Learning Through Data Intelligence. In R. A. Sottilare & J. Schwarz (Eds.), *Adaptive Instructional Systems* (pp. 174–187). Springer Nature Switzerland.