Navigating the Skies: Enhancing Military Helicopter Pilot Training Through Learning Analytics

Dirk Thijssen(✉), Pieter de Marez Oyens, and Jelke van der Pal

Royal Netherlands Aerospace Centre, Anthony Fokkerweg 2, 1059 CM Amsterdam, The Netherlands
{Dirk.Thijssen,Pieter.de.Marez.Oyens,Jelke.van.der.Pal}@nlr.nl

Abstract. Ensuring safety and proficiency in aviation requires effective training and maintenance of pilot skills. Pilots must maintain currency, implying regular flight, though some critical elements (e.g. emergency procedures) are rarely practiced in-flight. Simulator sessions are employed for safe practice and training. Simulators provide more opportunities for pilots to experience sophisticated training without safety risks and dependencies on for example weather and logistics. While human assessors currently evaluate pilot performance, a simulator can provide rich data for assessments, utilizing learning analytics for individual insights. Simulator data (and potentially aircraft data) is not regularly used yet in training to gain insights in performance. Furthermore, current training planning lacks consideration for individual needs, relying on fixed intervals and syllabi. This study aims to investigate how simulator data can be used to assess performance and if it provides valid statements about pilot performance. Thereafter, the data is used to develop a skill retention model accommodating personal differences. The research involves Chinook helicopter pilots from the Royal Netherlands Air Force, using simulator data to create tailored performance metrics for maneuvers. The ongoing study investigates the integration of simulator data into training assessments and aims to contribute valuable insights into pilot performance and skill retention. This paper presents the preliminary findings in the development of learning analytics within this context.

Keywords: Performance Assessment · Evidence-Based Training · Pilot Performance

1 Introduction

1.1 Related Work

Defence organizations are increasingly more focused on (operational) readiness, driven by a growing sense of urgency due to recent geo-political events, compelling them to be fully prepared for potential conflicts. In 2022, NATO leaders agreed on a new NATO Force Model [1, 21, 22]. This new model proposes a larger pool of high readiness forces

R. A. Sottilare and J. Schwarz (Eds.): HCII 2024, LNCS 14727, pp. 72–88, 2024.
https://doi.org/10.1007/978-3-031-60609-0_6

to improve NATO's ability to react in quick response. NATO members are commissioned to improve their operating methods in order to adhere to the new NATO Force Model. A critical component is the readiness of personnel. Readiness extends across every level of the organization, involving everyone from the military pilot, trained for specific goals and proficiency levels, to the commander, who requires thorough assessment reports and extensive data to make informed decisions regarding deployments and missions. Readiness can be subdivided into personnel, training, equipment on hand (i.e. amount) and equipment status (i.e. usability). Readiness can be assessed on an individual (e.g. pilot), team (e.g. unit) and collective level (e.g. joint forces) [2]. However, several divisions of readiness are available and can change per country and operational command.

Several studies from the RAND Corporation indicated that the current readiness assessment system falls short in effectively measuring the force's capacity to fulfill forthcoming mission requirements [2, 3, 23]. Investment in training assets is mentioned to help address the gaps in readiness assessment, specifically [2]:

- Mission Training through Distributed Simulation (MTDS)
- More simulators and new synthetic threat environments
- Aggregated force readiness measurement
- Adaptive, proficiency-driven training

The research investigates the utilization of data, particularly simulator data, for adaptive, proficiency-driven training. This field of research is known as learning analytics [4]. It involves the development of performance metrics to evaluate pilot performance. These metrics, in addition to instructor observations, help to outline trainee performance, leading to a more comprehensive assessment of readiness on an individual level.

Following this, the study explores the possibility of modeling performance over time based on the results of performance metric, which is called skill retention [5]. The integration of performance metrics and learning process modeling aims to optimize training by selecting and scheduling training tasks based on the individual needs of the trainee. This improves effectiveness of training and resource efficiency to meet the required proficiency level. Thereafter, results from a whole unit or force can be aggregated and analyzed from a commander's perspective, providing a comprehensive view of the readiness of the operating commands.

This paper presents the proposed method and initial development of performance metrics, along with preliminary results derived from the training session data. However, it does not delve into the modeling of skill retention. The performance metrics are crucial for modeling skill retention, and additional repeated measures are necessary. As data collection is ongoing, results will be published at a later stage. The research involves Chinook helicopter pilots from the Royal Netherlands Air Force (RNLAF), using simulator data to create tailored performance metrics for specific maneuvers. In the following section, we describe the Learning Analytics that may be applied to achieve this enhanced level of retention training and readiness.

1.2 Learning Analytics and Performance Metrics

Learning analytics is defined as the measurement, collection, analysis, and reporting of data on learners within their specific contexts. The primary objective is to comprehend and optimize the learning process and the learning environment [4].

The application of learning analytics further distinguishes itself based on the levels at which it operates. At one level, the focus is on individual (or team) learning processes, where performance is analyzed, providing valuable insights for self-evaluation, feedback, and performance assessment. Another level involves the analysis of data from large groups, offering valuable insights at the organizational level. Management can leverage these insights for improving learning trajectories and decision-making. In terms of military organizations, the management application is layered from managers of operational units responsible for training up to the commanders who decides on deployment and missions.

While the distinction between application at an individual or group level is well-known, another distinction is proposed between learning analytics within a training session and between multiple training sessions. Within the specific time frame of a single training session, learning analytics leverages data to provide insights into the performance of an individual. By extrapolating this functionality over multiple training sessions, aspects like skill retention can be addressed. While both types rely on similar data, distinct algorithms and analyses are required to effectively capture and assess performance trends over the course of time. For initial training, this could span weeks or months, while in continuation training, it may extend over months, years or even decades.

Types of Learning Analytics. A training session occurs with its own context, both foreseen (e.g. training scenario) and unforeseen (e.g. weather). Assessment of performance is highly dependent on this context and therefore learning analytics should be adjusted and fine-tuned for every context [6]. Military operators determine their actions based on the context, where a specific action may be correct in one context but improper in a slightly different one. The distinctions between contexts can involve subtle nuances but result in significant differences in desired behavior, particularly in critical situations. Regarding assessment, an instructor possesses the experience, knowledge and ability to recognize these differences based on their expertise and experience. An analytical model in learning analytics is less sensitive for the context and therefore it is important to provide sufficient context when interpreting the results. Models frequently tend to either overgeneralize or undergeneralize. Hence, it is crucial to determine the intended goal for using a performance metric and ensure its validity in that particular context. While there is currently no consistent framework for various types of learning analytics in military training, four distinct categories can be identified from the context of formal education: descriptive, diagnostic, predictive, and prescriptive analytics [4]. When applied to the training of a military operator, the definition slightly shifts from formal education.

Descriptive Analytics. Descriptive analytics offers insights into what takes place in the training scenario (e.g. visualizations of aircraft positioning relative to other objects). Descriptive analytics can be utilized in real-time and during an after-action review. Nevertheless, the insights presented by descriptive analytics still need interpretation

from the user since they solely describe what occurs and do not assess or assign value to the observed behavior. Descriptive analytics aids instructors in gaining awareness of aspects that may be challenging to observe directly. During an after-action review, it serves to replay events and reconstruct potential causes of both success and mistakes. Its primary application is within a training session.

Diagnostic Analytics. Diagnostic analytics offer insights into why something happened [4]. Unlike descriptive analytics, diagnostic analytics involves a degree of assessment. The conclusiveness of the assessment varies for each analytic. The least conclusive form involves indicating abnormal behavior, such as a pilot deviating from procedures. This indication does not categorically state correctness or incorrectness; instead, it highlights an event that is potentially interesting for reflection. An example of a diagnostic analytic with high conclusiveness is a model that grades the execution of a turn based on the prescribed procedures. These analytics can help an instructor focus attention on intriguing events or on competencies that are complex to assess (letting the analytics handle simpler competencies). For effective use, high conclusiveness must be granted validity. The examples are mainly useful within a training session.

A diagnostic analytic can be applied in a context beyond specific actions, such as assessing the execution of a turn. This involves the assessment of competencies. Vatral et al. [7] developed a method in which diagnostic analytics, grading specific actions or performance indicators, are utilized to grade competencies. The results of multiple specific metrics relevant to a competency are combined to generate a competency score. The competency framework can be multi-layered, where competencies from a lower level can also be combined to create a higher-level competency score. The methodology expresses the score of a metric between 0 and 1, and this score is further categorized into three states (below, at, or above expectation) for a metric or competency score. Bayesian modeling is employed to combine these scores into a higher-level competency score [7]. This Bayesian scoring method is applied within a training session. However, there is also potential to use such approach between training sessions for long-term competency assessment.

Another approach for long-term competency assessment between training sessions is the use of the Elo rating system [8]. The Elo rating system, originating from competitive gaming, quantifies the skill level of a trainee. Each participant is assigned a numerical rating for each competency, and after each training session, the ratings are adjusted based on the outcome. Every training task or scenario is also provided by a numerical rating per relevant competency, which is the complexity level. The updating mechanism of the Elo rating system involves a dynamic adjustment of players' skill ratings after each match. The extent of the adjustment is determined by the outcome of the training session and the difference between the trainee's skill rating and the complexity of the training tasks. A lower-rated trainee gains more from performing well in higher-rated training tasks, while a higher-rated trainee experiences a smaller rating increase when performing well against lower-rated training tasks. The difference between the skill rating of the trainee and the complexity level of the task is utilized to calculate a probability of the win chance, expressed as a percentage, indicating the likelihood of a good performance by the trainee given the task's complexity. The adjustment, determined by a logistic function, ensures a dynamic and continuous reflection of changes in trainee's skill levels over time. The

updating mechanism of the Elo rating system is visualized in Fig. 1. For a positive and effective learning curve, it is advisable to choose training tasks that align with the player's skill rating. This dynamic adjustment allows for the continuous assessment and monitoring of players' long-term competency, reflecting their evolving skill levels as they engage in more training sessions over time [8].

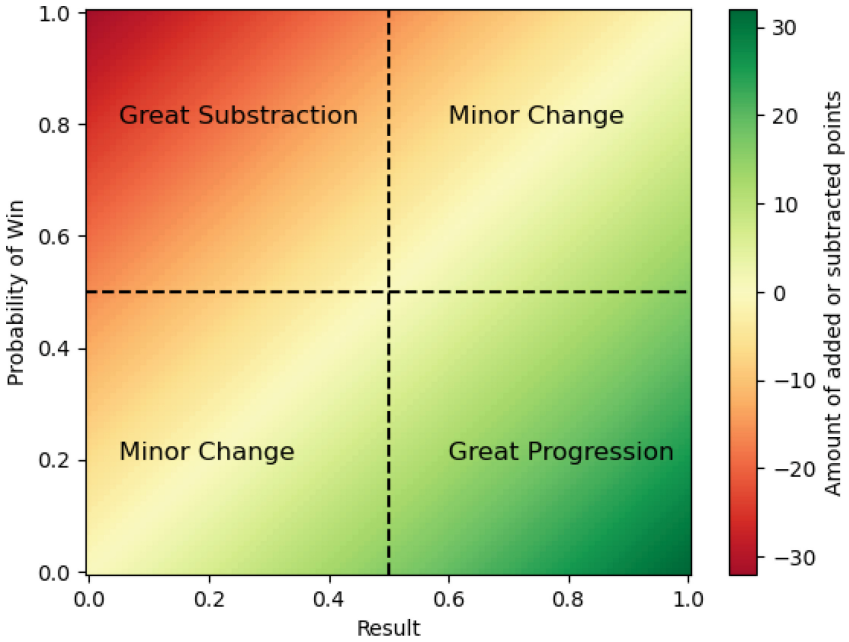

Fig. 1. A visualization of the updating mechanism of the Elo rating system.

Predictive Analytics. Predictive analytics are used to make estimates of future behavior. It combines historic data to identify patterns in the data and with the use of mathematical models and algorithms relationships between various variables can be captured in order to forecast trends [4]. These analytics are mainly useful between training sessions, for example in the prediction of skill retention. Anticipating when proficiency might diminish is advantageous for scheduling subsequent training sessions [9]. Various methodologies are employed to predict skill retention. The Predictive Performance Equation (PPE) adopts a theory-based approach, establishing a mathematical model based on: (1) the power law of learning, indicating that performance improves with practice; (2) the power law of forgetting, which suggests that performance declines over time since practice occurred; and (3) the spacing effect, emphasizing the benefits of distributed practice over time [9]. Another approach involves the utilization of machine learning algorithms, particularly those involving reinforcement learning and regression [10]. These approaches can also be integrated into hybrid models, where they complement each other [10].

Prescriptive Analytics. Prescriptive analytics provide recommendations and advice by utilizing historical data, such as scores, to identify training needs and relate them to various training options [4]. These analytics are frequently integrated into recommender systems, which can vary in their objectives and the techniques used to generate recommendations [11]. Recommendation systems can be applied across various time frames, as illustrated by the following examples. Firstly, a recommendation on very short notice involves the real-time adaptation of training scenarios based on the performance and mental state of the trainee. Performance, assessed by evaluator and/or other analytics, and mental state, determined by measures of brain activity, are used as inputs to adjust the scenario's complexity. For instance, a fighter pilot may encounter more or fewer adversary aircraft based on their performance [12]. Secondly, historical performance data can be utilized to construct the next scenario in a training. In the Pilot Training Next research program, historical assessment data and the scenarios constructed by instructors based on that data are provided to an Artificial Intelligence (AI) model. The model suggests the next training scenario by emulating the instructor's scenario construction process, considering the individual needs of each trainee, derived from the data [13]. The third example occurs on a similar timeline to the Pilot Training Next example. By employing the Elo rating system, competency scores are generated and updated over time. Competency scores that lag behind in development are identified, and subsequent training activities are recommended to address those underperforming competencies at a complexity level that matches the trainee's skill level. To facilitate this, an Experience Index is formulated, detailing which competencies are trained by each training task and at what complexity level [8].

Measures in Military Exercises. The utilization of analytics or measures is not entirely novel in the military domain. During military exercises, measures are frequently employed to obtain a more objective assessment of task execution. Two types of measures are defined: Measures of Effectiveness (MoE) and Measures of Performance (MoP). An MoE describes the desired outcomes, such as the mission objective, while an MoP characterizes the performance of an action irrespective of the overarching mission goal. These measures are established beforehand and evaluated afterward. Evaluation can be based on operator observations and opinions during debriefs, but outcomes can also be determined based on data [14]. The types of analytics mentioned in the previous section, primarily descriptive and diagnostic analytics, could also serve as input for determining the outcomes for MoE and MoP.

2 Methodology for Performance Metrics and Competency Scores

For modelling and optimizing retention training, our proposed method makes use of the methodology and results of Vatral et al. [7], which described the Generalized Intelligent Framework for Tutoring (GIFT). This research uses the Generalized Intelligent Framework for Tutoring (GIFT). GIFT is pioneering in the development of AI-based Augmented and Intelligent Tutoring Systems (IAAR-ITS) environments [7, 19, 20]. The research employs a competency-based approach, breaking down complex tasks into competencies and subtasks through cognitive task analysis, resulting in a layered competency profile. Each subsequent layer in the hierarchy further decomposes the complex

task, specifying subtasks until the level of observable behavior and actions is reached. The bottom layer of the profile is made measurable through performance metrics [7, 20].

Subsequently, performance metrics developed for the bottom layer of the profile, representing observable behavior. The exact method for developing the analytics is not generic, as creating analysis models is context-dependent. The performance metrics are described in detail in Sect. 3.4. The layered competency profile allows scores at the bottom to be aggregated to higher layers. Figure 2 depicts the interactions between the different layers precisely, illustrating how a subordinate component can contribute to multiple superior components. The multi-layered competency profile that is shown is only a placeholder and will be defined at a later stage. A distinct visualization can be made for each maneuver or task [7, 20]. The orange boxes indicate that assessments scores of instructors can also be included in this method. This was not part of the initial methodology [7, 20].

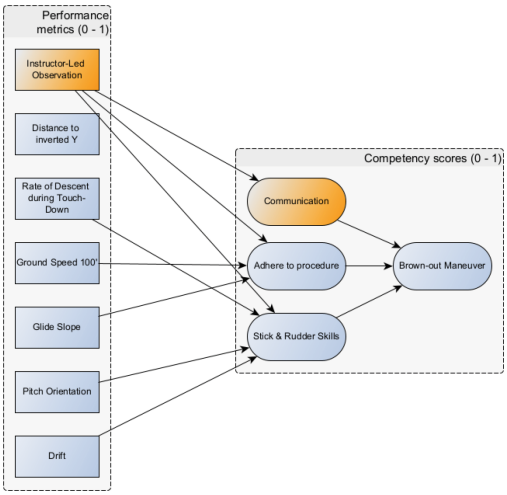

Fig. 2. An overview of the interaction between the performance metrics (left box) and the multi-layered competency profile (right box) for the brown-out maneuver.

A performance metric can thus contribute to multiple competencies [7, 20]. Watz et al. describe a criterion-based approach for developing learning analytics, defining rules and cut-off points with experts to convert measurements into assessments. An effective analysis model can be created with sufficient data and a well-defined concept of success. Watz et al. also investigate aggregating underlying components to higher-level ones, combining various measurements, both subjective and objective, into a performance score through a weighted average. Initial weights are assigned by experts and can later be refined using machine learning techniques [6]. However, Vatral et al. use a different method to aggregate scores of underlying components to higher-level ones [7, 20].

The result of a performance metric at the lowest level of the framework is a continuous number between zero and one, with three labels specified: (1) below expectation for a

score below 0.4, (2) at expectation for a score between 0.4 and 0.9, and (3) above expectation for a score above 0.9. The continuous number can be based on a percentage or score. A continuous score between zero and one is used because Bayesian network models are employed to generate assessments for higher levels in the hierarchy. Bayesian networks are probabilistic models, used for predictions, representing relationships between variables. They consist of nodes representing variables and directed arrows indicating causal relationships between variables. Each node has a probability distribution representing the likelihood of each possible outcome of that variable, given the values of the variables it is linked to. These probability distributions are updated using Bayesian inference, using new data to calculate the likelihood of each possible explanation and determining which explanation is most likely. For each competency, the probability of the state (i.e. below-, at- and above-expectation) is calculated [7]. The Bayesian network model in this research consists of four models:

1. Three state competency model: the likelihood of a competence status given the result of a performance metric (i.e. below-, at- or above-expectation).
2. Transition model: determining the likelihood of a competence status changing after a training event (e.g., from below- to at-expectation).
3. Conditional model: assessing the likelihood of the higher-level competency having the same status as the underlying competencies.
4. Prior model: evaluating the probability of the team having a certain status based on prior knowledge and experience.

The method enables personalized training sessions by updating predictions within each session, optimizing training duration to maximize learning gains while minimizing diminishing returns. Additionally, it integrates into a larger ecosystem for evidence-based, data-driven trainee assessments across various scenarios and environments, ensuring consistent and reliable trend analysis regardless of training location or instructor.

3 Method

3.1 Experiment Design and Data Collection

Experiment Design. The experiment involved two groups of pilots: one conducting sorties every three months and the other every six months. The data collection period spans two years. These time intervals were selected to allow sufficient time for skill decay to occur. The discrepancy in time between the two conditions is implemented to explore differences in pilot performance after a specified duration. It is presumed that other flight experiences may impact the skill retention process. Therefore, in consultation with Subject Matter Experts (SME), three maneuvers are selected that are infrequently performed, or alternative procedures are employed to minimize the influence of other flight experiences. Additionally, the total number of flight hours and flight hours over the past year are recorded to assess whether these factors influence performance.

The three maneuvers performed are: brown-out landings in sandy environments, autorotation in case of engine failure, and instrument flying with cockpit instruments only. Variations were applied to the scenarios for each maneuver, but deviations from the proposed schedule resulted in scenarios being flown at different times or not at all.

The brown-out scenarios are the same for every training session. Four attempts are made in each training session. Participants are required to land to helicopter as close to a certain point as possible. This point is between two lamps within a formation of lamps, called inverted Y. Therefore, the brown-out maneuver only differs in timing (refresh training after three or six months).

Five scenarios are developed for the autorotation maneuver, conducted in an open field which provides plenty of landing spots. The scenarios vary in the location where they are conducted. Each training session involves one of these five scenarios. Additionally, two more complex scenarios are created, set above forests where fewer suitable landing spots are present, challenging pilots to make rapid and appropriate decisions. One complex scenario is included in training sessions four and eight for the three-month condition, and sessions two and four for the six-month condition. The simple scenario is repeated twice within a training session, while the complex scenario is executed only once.

Instrument flying is conducted in both a simple and complex scenario during every training session. The simple scenario alternates between two options for each session, while there are eight unique scenarios available for the complex scenario. Each type of scenario is performed once in each session. Besides that, the instrument flying differs in timing (refresh training after three or six months).

Data Collection. The data collection consists of saving simulator data and a questionnaire filled in by the participant. The simulator exports data in the DIS-protocol, which provides location, orientation and several speed parameters. The DIS-data is eventually converted to a CSV-file (see Sect. 3.4).

The questionnaire is filled out by the trainee after every training session. The first part is filled out before the training sessions starts. The following information is requested before the training session:

- The total amount of flight hours (rough approximation);
- The percentage of the total amount of flight hours on the CH-47F MYII CAAS (rough approximation);
- The total amount of flight hours in the last twelve months (rough approximation);
- Whether or not the participant is a staff pilot;
- Karolinska Sleepiness Scale (KSS) [17];
- A rating on self-efficacy on each maneuver between one and ten.

The following information is requested after the training session:

- A rating of perceived cognitive load on each maneuver between one and seven;
- A rating of perceived success on each maneuver between one and seven;
- A question whether the perceived cognitive load and/or success differed between the scenarios within a training session (including space to elaborate on their experience).

3.2 Apparatus

The research is conducted using the Transportable Flight Proficiency Simulator (TFPS). This high-fidelity simulator is utilized by CH-47 pilots of the RNLAF primarily for training (emergency) procedures. Efforts are underway to integrate the TFPS with other simulators, such as a second TFPS, AH-64 simulator, and Rear Crew Trainers (RCT). Consequently, tactical training will also be incorporated into the simulator's usage [15] (Fig. 3).

Fig. 3. The Transportable Flight Proficiency Simulator (TFPS) built by NAVAIR [16].

The pilots undertake a specially designed sortie for the research, featuring three maneuvers: brown-out, autorotation, and instrument flying. During the brown-out maneuver, pilots perform a landing in a sandy environment where the final phase of the landing lacks outside visuals due to swirling sand, potentially causing disorientation. While standard procedures typically involve using the autopilot, this research employs a more manual procedure. An autorotation, an emergency maneuver, is executed when the engines malfunction. Pilots initiate a steep dive to convert altitude into kinetic energy for the rotor head. As the ground approaches, this kinetic energy is utilized to reduce vertical speed and safely land the helicopter. Instrument flying restricts pilots to utilizing only the instruments within the cockpit, without relying on visual cues from outside. This practice is particularly valuable in low-visibility conditions such as nighttime or foggy weather. Arrival routes are flown using only cockpit instruments in this research setting.

3.3 Participants

Participant in this research were CH-47 pilots of the RNLAF. The complete population of active CH-47 pilots in the RNLAF is approximately sixty pilots, from which twenty

pilots are only flying minimally due to other job responsibilities. In the present research 26 pilots are selected to participate. The participants are equally divided between the two conditions. As of the current report, eighteen training sessions have been completed. Specifically, one participant completed two sessions, while another completed three. Thirteen participants each completed one session. Unfortunately, data collection failed during three sessions.

3.4 Data Analysis

This paragraph describes the various activities performed for data analysis, all of which are conducted using Python.

DIS-Converter. The DIS protocol is primarily designed for connecting simulators and does not inherently support the development of learning analytics. To enable analysis, a software module was developed to convert DIS data into a CSV file format suitable for analysis. This module extracts key information such as timestamp, coordinates (i.e. x, y, z), orientation (i.e. pitch, roll, yaw), and various speed variables from each DIS message, organizing them into rows in the CSV file. This CSV file containing the extracted variables can then be utilized for the development of learning analytics.

Dashboard for Labeling. In each sortie, the three maneuvers are performed multiple times (i.e. attempt), and this data is recorded in one session per person. To create separate CSV files for each attempt, data extraction and labeling is necessary. While exploring automatic data labeling approaches, it was deemed that the effort to develop such software outweighed the time it would save. Instead, a dashboard was developed to visualize the altitude over time for each training session. The dashboard is depicted in Fig. 4. This visualization enabled the user to identify the different maneuvers and attempts based on their knowledge of the experiment. Users could then manually add starting and ending points to the timeline on the dashboard. After data collection, the recordings were processed through the dashboard, and the starting and ending points were used

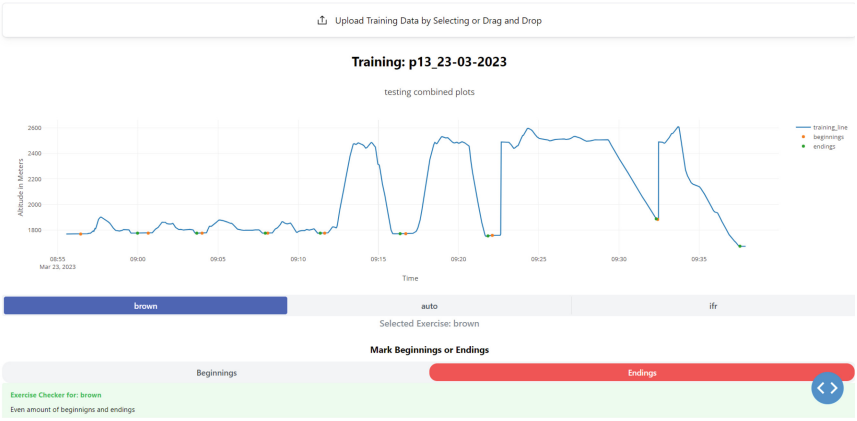

Fig. 4. Dashboard developed for data labeling, displaying time (x-axis) and altitude (y-axis) with starting (orange) and end points (green) to distinguish maneuvers. (Color figure online)

to generate distinct CSV files for each maneuver. Although the dashboard provided a solution for data labeling in a specific format with a specific goal, the lack of integrated data labeling tools or solutions in training devices remains a recurring issue [18].

Learning Analytics. Performance metrics were identified partly by consulting flight and training manuals, as they detail the execution of maneuvers and required speeds, among other factors. Also, the available data types, such as position and orientation of the helicopter, are considered. The design of the metrics was further refined through consultation with a simulator instructor, a qualified CH-47 pilot.

The performance metrics are listed below in Table 1. Each metric is accompanied by its relevance to specific maneuvers and a brief description of how it is computed. The performance metrics yield continuous scores, such as the distance in meters for the metric "distance to inverted Y." These scores are then normalized to a range between zero and one using predefined cut-off points, such as the distance at which a landing is deemed out-of-bounds. Subsequently, these normalized scores can be categorized as below, at, or above expectation levels. However, specific cut-off points are not disclosed due to classification reasons.

Table 1. The performance metrics with its relevance to the maneuver and a basic description on how they are computed.

Performance metric	Relevant maneuver	Computation
Distance to inverted Y	Brown-out	The pilots must land the helicopter at a specific location (the inverted y). The distance is calculated between the inverted y and the location the helicopter came to a standstill
Rate of Descent during Touchdown	Brown-out Autorotation	The moment of touchdown is determined. Thereafter, the vertical speed at that moment is compared to the cut-off points
Ground Speed 100'	Brown-out Autorotation	During landing the ground speed should decrease. The ground speed as reviewed at 100 ft. and 50 ft. in relation to the cut-off points
Glide Slope	Brown-out Autorotation	The glide slope in the last ten timesteps is compared to the cut-off points

(*continued*)

Table 1. (*continued*)

Performance metric	Relevant maneuver	Computation
Pitch Orientation	Brown-out Autorotation	The pitch in the last ten timesteps is compared to the cut-off points
Drift	Brown-out Autorotation	The roll and yaw in the last ten timesteps are compared to the cut-off points
Decision Making - Distance versus Time	Autorotation	The pilot must decide based on available landing sites between maximizing flight time or covering more distance with remaining kinetic energy after engine failure. Each option corresponds to a preferred speed range, determined by specific cut-off points
Correlation actual and suggested flight path of the approach	IFR	Every approach has a suggested flight path that must be followed. The correlation between the actual and suggested flight path is calculated indicating the adherence to the approach
Deviation from localizers	IFR	In an approach the pilot must adhere to a maximum and minimum altitude and wide based on so-called localizers. The absolute margins decrease when the runway is approached. The deviation/adherence from these minimum and maximum are calculated at several moments in the approach

4 Results

The paper presents preliminary results, with ongoing data collection and analysis. Although the analysis is still in progress, the initial version of three performance metrics, applied to the brown-out maneuver, is presented here. The algorithms have been applied to a dataset comprising fifteen training sessions. However, data from three sessions are corrupted, and efforts are underway to address this issue. Additionally, not all training sessions completed the planned four brown-out attempts. The preliminary results are presented in Table 2.

The results provide insight in the item discrimination of the performance metrics. Item discrimination is the measure of how well an item (e.g. test question, performance metric) can distinguish between those who perform well and those who do not. The

Table 2. Preliminary results for three brown-out maneuver performance metrics: scores categorized as below (0), at (1), and above (2) expectation. Mean scores are presented due to balanced category intervals, alongside standard deviation (SD) and response count (N).

		Mean	SD	N
Distance to Inverted Y	Attempt 1	0.25	0.62	12
	Attempt 2	0.33	0.65	12
	Attempt 3	0.11	0.33	9
	Attempt 4	0.63	0.92	8
Rate of Descent during Touchdown	Attempt 1	1.67	0.78	12
	Attempt 2	1.75	0.62	12
	Attempt 3	1.33	0.87	9
	Attempt 4	1.13	0.83	8
Pitch Orientation	Attempt 1	0.58	0.90	12
	Attempt 2	1.70	0.72	12
	Attempt 3	1.22	0.83	9
	Attempt 4	1.00	0.76	8

metric of 'distance to inverted Y' seems to have a poor item discrimination based on the low mean score. Approximately 80% of the attempts were labelled below expectation. This is also indicated by the relatively low standard deviation compared to the two other performance metrics.

5 Discussion

The preliminary results are limited; thus, it is premature to draw conclusions regarding the validity of the performance metrics or the pilots' performance. Nonetheless, this phase has provided valuable insights into the development and implementation of learning analytics in helicopter pilot training.

5.1 Design and Development Process of Performance Metrics

During development and computation of the performance metrics several challenges occurred. First, it was challenging to create a consistent algorithm to handle unexpected data, such as missing attempts or prematurely ended maneuvers (e.g., go-arounds). Additionally, obtaining conditional information necessary for computing performance metrics proved challenging. For instance, determining the exact moment of touchdown of the helicopter must be derived from simulator data, as there was no explicit label or variable provided. Establishing a consistent method to determine touchdown and similar information posed difficulties. Therefore, this version represents an initial attempt and may require refinement in subsequent iterations. These challenges align with the data labeling issue outlined by Bessey et al. [18]. Frequently, datasets lack standardized labeling,

such as consistent start and stop times or event identification. Addressing this labeling problem enhances the data's utility and reduces the analysis workload. Possible solutions may include real-time labeling tools for simulator operators or classifier algorithms for automated data labeling, such as identifying maneuvers or helicopter touchdowns.

5.2 Item Discrimination and Validity

The preliminary results showed that item discrimination could be a potential problem regarding the validity of the performance metrics. In the continuation of this research it can be beneficial to statistically investigate the item discrimination, for example with the item-response theory.

Adjustments could enhance the usefulness of the performance metrics, such as improving item discrimination and validity. One planned adjustment is transitioning the outcome of the metric from a categorical to a continuous variable. For instance, consider the metric "distance to inverted Y," which measures the distance between the touchdown point and the intended landing spot. Currently, true/false statements check whether the distance belongs to the criteria of below, at and above expectation, resulting in a categorical variable. The true/false-statements reduces the variability massively in comparison to a continuous variable. To generate a continuous variable, the distance could be related to a minimum distance indicating perfect execution and a maximum distance indicating poor performance. This approach could yield a continuous variable ranging from zero to one, with values then categorized as below, at, or above expectation, aligning with suggestions by Vartral et al. (e.g. at expectation for a score between 0.4 and 0.9) [7, 20].

Furthermore, a review of the cut-off points could be considered. If a significant number of participants' attempts are consistently classified as below expectation for example, adjusting the cut-off points to enhance item discrimination might be advantageous. However, from an operational standpoint, altering established cut-off points could be undesirable. Doing so could potentially lead to either accommodating incompetent pilots if requirements are relaxed or losing competent pilots if requirements are made more stringent.

Additionally, it is important to assess whether the performance metrics accurately label performance compared to assessments made by instructors. In future phases of this research, training sessions will be organized wherein trainees are evaluated using the performance metrics alone, by an instructor, and by an instructor aided by the performance metrics. Examining the correlation between these three assessments can provide insights into the validity of the performance metrics.

5.3 Limitations and Future Research

One limitation of the research is the inconsistent attendance of participants, largely due to high operational demand, resulting in deviations from the proposed training schedule. This variability in the number and timing of training sessions significantly diverges from the initial plan. It is anticipated that this deviation may impact the statistical significance within tests and could potentially hinder the development of skill retention models.

One limitation of the research stems from the use of univariate analysis rather than multivariate analysis [7, 20]. The performance metrics in this study rely solely on DIS-data, which, while valuable, may be limited in capturing the full scope of activities. Future research should explore incorporating additional data types to enhance analysis.

Considerable knowledge and experience are being accumulated in regard to learning analytics, providing insights for developing new types and applications of performance metrics. It is crucial to establish a modular and expandable environment for these metrics, addressing aspects such as model and algorithm storage, version management, and the incorporation of feedback to refine results from previous iterations. Future research should focus on investigating the architecture of such an environment to support ongoing advancements in the field.

Acknowledgments. This study is part of the Adaptive Learning Ecosystem program, funded by the Dutch Ministry of Defense (grant number L2201).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wolting, J.: NAVO verandert huidige Force Model. Defensie Magazine - Landmacht 07 Artikel 8 (2022). https://magazines.defensie.nl/landmacht/2022/07/08_nieuwe-nato-force-model
2. Emmi, Y., et al.: How Training Infrastructure Can Improve Assessments of Air Force Readiness. RAND Corporation, Santa Monica (2023). https://www.rand.org/pubs/research_briefs/RBA992-1.html
3. Emmi, Y., et al.: Air Force Readiness Assessment: How Training Infrastructure Can Provide Better Information for Decisionmaking. RAND Corporation, Santa Monica (2023). https://www.rand.org/pubs/research_reports/RRA992-2.html
4. Society for Learning Analytics Research. https://www.solaresearch.org/about/what-is-learning-analytics/
5. Chittaro, L., Van der Pal, J., Oprins, E., Van Puyvelde, M., Taylor, H., Rankin, K.: Skill Fade and Competence Retention: A Contemporary Review. Brussels: North Atlantic Treaty Organization - Science and Technology Organization (2023)
6. Watz, E., Neubauer, P., Kegley, J., Bennet, W.: Managing learning and tracking performance across multiple mission sets. In: I/ITSEC (2018)
7. Vatral, C., Biswas, G., Naveeduddin, M., Goldberg, B.: Automated assessment of team performance using multimodal Bayesian learning analytics. In: I/ITSEC (2022)
8. Thijssen, D., Bosma, R.: Recommendation system in an integrated digital training environment for the 5th generation air force. In: I/ITSEC (2022)
9. Walsh, M.M., Gluck, K.A., Gunzelmann, G., Jastrzembski, T., Krusmark, M.: Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. Cogn. Sci. **42**(S3), 644–691 (2018). https://doi.org/10.1111/cogs.12602
10. Sense, F., et al.: Combining cognitive and machine learning models to mine CPR training histories for personalized predictions. Int. Educ. Data Min. Soc. (2021)
11. Uddin, I., Imran, A., Muhammad, K., Fayyaz, N., Sajjad, M.: A systematic mapping review on MOOC. IEEE Access (2021)

12. Tillema, G., Roza, M.: Data-driven and personalized training as a service infrastructure & technologies. In: I/ITSEC 2023, Orlando (2023)
13. Forrest, N.: Conceptualization and Application of Deep Learning and Applied Statistics for Flight Plan Recommendation. Air Force Institute of Technology, Ohio (2020)
14. Civil-Military Cooperation Centre of Excellence. Measures of effectiveness (MoE) and measures of performance (MoP). CIMIC Handbook (2020). https://www.handbook.cimic-coe.org/
15. Royal Netherlands Aerospace Centre. Case: Multi-Ship Multi-Type Helicopter Simulation Training Capability (n.d.). https://www.nlr.org/case/case-multi-ship/
16. Berry, T.: Transportable Flight Proficiency Simulator [Photograph]. Vertical Magazine: Bron (2021). https://verticalmag.com/press-releases/u-s-and-international-crew-members-train-on-flight-simulators-in-germany/
17. Akerstedt, T., Gillberg, M.: Subjective and objective sleepiness in the active individual. Int. J. Neurosci. **52**, 29–37 (1990)
18. Bessey, A., Waggenspack, L., Schreiber, B., Bennet, W., Jr.: Tackling the human performance data problem: a case for standardization. In: I/ITSEC (2022)
19. Goldberg, B., et al.: Forging competency and proficiency through the synthetic training environment with an experiential learning for readiness strategy. In: I/ITSEC (2021)
20. Vatral, C., Naveeduddin, M., Biswas, G., Goldberg, B.: GIFT external assessment engine for analyzing individual and team performance for dismounted battle drills. In: Ninth Annual GIFT Users Symposium (2021)
21. Twigt, A.: Task Force Defensienota 22 wil samen het verschil maken. De Vliegende Hollander (2023). https://magazines.defensie.nl/vliegendehollander/2023/09/06_successen-van-tf-22_slot
22. Monaghan, S., Wall, C., Morcos, P.: What Happened at NATO's Madrid Summit? [Critical Questions] (2022). http://tinyurl.com/444wthkm
23. Walsh, M., Taylor, W.W., Ausink, J.A.: Independent Review and Assessment of the Air Force Ready Aircrew Program: A Description of the Model Used for Sensitivity Analysis. RAND Corporation, Santa Monica (2019)