Conceptualization and Application of Deep Learning and Applied Statistics for Flight Plan Recommendation

THESIS

Nicholas C. Forrest, Second Lieutenant, USAF

AFIT-ENS-MS-20-M-147

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-20-M-147

CONCEPTUALIZATION AND APPLICATION OF DEEP LEARNING AND APPLIED STATISTICS FOR FLIGHT PLAN RECOMMENDATION

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Nicholas C. Forrest, B.S.

Second Lieutenant, USAF

March 2020

AFIT-ENS-MS-20-M-147

CONCEPTUALIZATION AND APPLICATION OF DEEP LEARNING AND APPLIED STATISTICS FOR FLIGHT PLAN RECOMMENDATION

THESIS

Nicholas C. Forrest, B.S.
Second Lieutenant, USAF

Committee Membership:

Dr. Raymond R. Hill
Chair

Jason R. Anderson, Lt Col, Ph.D.
Member

Phillip R. Jenkins, Capt, Ph.D.
Member

AFIT-ENS-MS-20-M-147

Abstract

The Air Forces Pilot Training Next (PTN) program seeks a more efficient pilot training environment emphasizing the use of virtual reality flight simulators alongside periodic real aircraft experience. The PTN program wants to accelerate the training pace and progress in undergraduate pilot training compared to traditional undergraduate pilot training. Currently, instructor pilots spend excessive time planning and scheduling flights. This research focuses on methods to auto-generate the planning of in-flight events using hybrid filtering and deep learning techniques. The resulting approach captures temporal trends of user-specific and program-wide student performance to recommend a feasible set of graded flight events for evaluation in a student's next training exercise to improve their progress toward fully qualified status.

Contents

Page

List of Figures

Figure		Page

Figure | Page

Figure Page

Figure Page

Figure Page

Figure Page

List of Tables

CONCEPTUALIZATION AND APPLICATION OF DEEP LEARNING AND APPLIED STATISTICS FOR FLIGHT PLAN RECOMMENDATION

I. Introduction

1.1 Background

The National Defense Strategy of The United States of America establishes that the enduring mission of the Department of Defense is providing military forces capable of deterring war and protecting the security of the United States and its allies [1]. Each branch of the Department of Defense publishes an official operations business plan to ensure that their operations are properly aligned with the mission established in the National Defense Strategy and directed towards meeting mission requirements. For the Department of the Air Force, maintaining overall force readiness has been identified as an essential aspect of effectively executing this mission despite a national pilot shortage. In a hearing before the Subcommittee on military personnel of the Committee on Armed Services of the House of Representatives, Lieutenant General Grosso, United States Air Force Deputy Chief of Staff for Manpower and Personnel Services, emphasized the true extent of the Air Force pilot shortage. Grosso explained that funding limitations and an increase in demand for Air Force pilots in the commercial airline industry have had a negative effect on force capacity and mission capability in anticipation of a potential full-scale fight against adversaries. At the end of fiscal year 2016, the Air Force's total force structure was 1,555 pilots short of requirements needed meet national security demands. Lieutenant General Grosso noted that the commercial airline industry currently hires over 4,000 pilots annually, offering increasingly higher salaries. Grosso has identified the need for pilot production and the Air Force's progressive focus on developing creative, agile solutions to meet pilot demands [2].

The impacts of the Air Force pilot shortage on force readiness have spread throughout the Air Force. In a 2018 hearing, Air Force Secretary Heather Wilson reconfirmed that the United States operates "in a more competitive and dangerous international security environment than we have experienced in decades. So the restoration of the force, the restoration of the readiness of the force, to win any fight, any time has to be job one for all of us"[3]. Wilson emphasized that the first evidence of readiness recovery is through force size. While production and retention of all career fields are essential, pilot production is at the forefront. The Air Force managed to train 1,160 pilots in FY-17 and expects to reach a steady-state of 1,500 pilots per year from FY-22 onward [3].

Pilot production stems from the ability to effectively and efficiently train pilots. Increasing qualified pilot production requires the Air Force to provide sufficient time and realistic training environments for Airmen to develop at accelerated paces. Currently, Undergraduate Pilot Training (UPT) operates in a three phase system that spans about one year. The first phase introduces students to basic aircraft control and flying with instruments in an academic environment. The T-6 aircraft is the first aircraft that students work with. Outside of the classroom, students conduct T-6 simulator missions to gain flight experience.

At the end of the first phase, students undergo a simulator-conducted check ride to assess basic aircraft control abilities and their flying knowledge. In phase two, daily evaluations are scheduled, conducted, and reviewed by an Instructor Pilot (IP), introducing students to legitimate flight hours in the T-6 aircraft. Phase two begins with a series of basic flight events and transitions into training blocks focused on formation and navigation events. When a student performs a flight event during a training exercise, the IP grades their performance as unsatisfactory, fair, good, or excellent. Each of these grades corresponds to a score of 1 to 4, respectively. Students continue to progress through flight events until deemed proficient in all necessary flight events. Each possible flight event taught in undergraduate pilot training has a required score to represent proficiency, called a maneuver item file (MIF) score. The MIF scores for the current set of possible events are all 3 or 4, indicating that the student must receive a good or excellent score for each flight event to be deemed proficient.

Upon completion of the formation and navigation training, students begin moving into more specified training tracks in either the T-38 or T-1 aircraft. When assigned to the new aircraft, students enter into phase three of their training. Phase three consists of similar training schedules as phase two but in a particular aircraft. If students are deemed fully proficient within an aircraft, they will graduate from UPT at the end of phase three and move on to their next duty assignment destined to operate in their newly assigned aircraft [4]. The UPT process is designed to thoroughly expose students to a vast range of flying skills and training environments.

The pilot shortage introduces pressure on the pilot training program to produce more pilots at an accelerated pace. Increasing pilot production via the traditional pilot training pipeline requires an increase in funding, equipment, and most importantly manpower resources that are currently unobtainable given total force personnel shortages. The current Air Force pilot training pipeline already suffers from operational delays due to the lack of resources, introducing months of wait time between when Airmen are assigned to attend pilot training and their official training start dates [5]. An inability to increase resources may result in less favorable alternative approaches to increase pilot production. One alternative method involves decreasing pilot training graduation requirements to the bare minimum skill set, thereby producing more minimally qualified pilots at an accelerated pace. However, executing similar alternatives only adds to a list of inconsistencies within the UPT program.

Secretary Wilson noted that the Air Force must move past traditional methods and implement virtual and constructive training in order to meet current and future force readiness demands, indicating that "sometimes now you can do more in a simulation that you can do actually up in the air"[3]. The Air Force has initiated the Pilot Training Next (PTN) program in Austin, Texas as an attempt to combat the current pilot shortage without sacrificing training expectations. PTN aims to provide a more personalized pilot training experience through the emphasized use of virtual reality flight simulators alongside periodic real aircraft experience to progress pilot training students through their pipeline in less time than it takes to complete traditional UPT.

PTN supports four primary factors that make it a plausible alternative to the traditional UPT structure: immersive technology; unlimited simulator availability; adoption of a free rein, non linear syllabus; and experience in a high risk, low reward environment. Immersive virtual reality training allows for students to receive satisfactory training time at reduced resource and maintenance costs compared to training in a real aircraft. Virtual reality training also reduces overall strain and time spent preparing and maintaining aircraft for flight. Traditional UPT only provides students the opportunity to get flight or simulation training during scheduled hours, but PTN provides students access to flight simulators at all times of the day and night. Students have access to larger, more realistic simulators in the office and smaller simulators provided at home. Unlimited access to flight simulators allows students to continue practicing beyond daily duty hours. Emphasizing the use of flight simulators gives students more access to training regardless of the weather, time of day, or aircraft availability. PTN does not follow a predefined syllabus. Instead, the training focuses on particular pilot training student competency, giving students the opportunity to advance in training at their own pace.

Unlike its traditional counterpart, PTN allows students to perform multiple training exercises in a single day. A training exercise consists of either a simulated or real flight where a student is evaluated on a series of flight skills. Each skill subject to evaluation is defined as a flight event. The culmination of all training exercises completed by a single student throughout their pilot training experience is defined as that student's training campaign. The use of virtual reality also submerses students into an environment where they can take risks and make mistakes without costly consequences. Such an environment encourages students to try new things and learn from them. The environment mitigates concerns of over accelerated advancement because lives are not at stake when trying new events in a simulated environment. Transitioning to more personalized training methods will result in higher variance for graduation dates, but provides the opportunity for students to progress through training faster than in an environment with a set syllabus. The first PTN graduating class has shown that it is possible to properly prepare Air Force pilots in about six months rather than the twelve-month schedule that UPT currently adheres to.

PTN shows potential to help solve the pilot shortage, but the current training evaluation structure across all pilot training methods allows for various forms of subjectivity to be introduced. Although standardized evaluation criteria exist, individual performances are subject to an individual IP interpretation resulting in flight-to-flight grading subjectivity. Other than instructor-to-instructor inconsistencies, the current grading system suffers from subjective inconsistencies across the entirety of the training. Performances on a given event that earn higher scores during the early stages of training are often not sufficient enough to receive the same score on the same event later in the program. Scoring inconsistencies led to inconsistencies in determining the overall proficiency of individual students. Additional procedural requirements, such as filing official paperwork for students that show regression in training, also introduce incentives for IPs to continue to award higher scores as students progress through training. This process has the potential to allow students to progress through training without achieving the proper training.

PTN and AFWERX-Austin are conducting a joint initiative to produce and integrate an automated IP system, known as the AutoGradebook, within the PTN program to eliminate evaluation subjectivity and inconsistencies. The AutoGradebook design consists of four primary components: an event recognition and grading component, a feedback component, an overall scheduling tool based on outside influences, and a next flight recommendation component. During a training exercise, students are graded on their individual ability to properly complete flight events. Scores are assigned in accordance with performing the correct event at the correct time under proper specifications such as speed and timing. Real-time feedback is provided based on student performance. A scheduling tool takes into account outside forces such as weather and physiological factors to schedule training. Recommendations are then provided in accordance with a student's culminating scores across all events and overall progression through training. Currently, IPs spend hours manually performing these administrative operations. Performing administrative tasks restricts the amount of time allocated for actual training and instructing. An automated instructor recommender system allows for IPs to focus on student development rather than administrative work.

Aside from eliminating subjectivity from the evaluation process, introducing the AutoGradebook allows for faster and more accessible training feedback. Mobile simulator units equipped with AutoGradebook technology dispatched to Service Academies or Reserve Officer Training Corps (ROTC) detachments present an opportunity to provide potential pilot training students with flight experience before ever arriving at pilot training. PTN and AFWERX leadership are confident that the AutoGradebook can introduce a collective, automated IP opinion that objectifies the evaluation element of pilot training to ensure a consistent skill standard while accelerating pilot production rates.

1.2 Overview

This study conceptualizes a recommendation system for upcoming training exercises during PTN based on recorded evaluation data from the previous PTN graduating classes.

1.3 Problem Statement

Every day, IPs spend hours reviewing, organizing and developing flight plans for pilot training students. Committing valuable time to various administrative tasks required to progress a student through training prevents IPs from focusing on personalized instruction and the advancement of the students. This study applies analytics to conceptualize and develop a recommendation system that effectively provides an IP with flight event suggestions for a student's next flight plan. The algorithm automatically populates a list of recommended flight events well suited for a student's next training exercise. Recommendations are generated based on each student's prior flight plans, grades received in prior training exercises, and overall progression compared to other pilot training students. Success is quantified by the number of similar recommendations the algorithm provides for a given flight plan compared to a flight plan generated from an IP.

1.4 Research Scope

This research focuses primarily on the conceptualization of a recommendation system that can effectively provide an IP with recommendations for training events that a pilot training student should perform in the next step of their training. IP, student, and flight dates are the primary factors influencing the segmentation of recommendations.

PTN leadership provided raw data collected from their first pilot training class. The data set consists of the scores received on every event performed during each training exercise for the nineteen students in the original PTN class. Every flight event does not have to be performed for each flight. Students are identified by an identification number. Data provided does not include personally identifying information. Calendar dates when individual training exercises were performed were recorded. All data provided are assumed as extensive and correct. The order of each student's training exercises is assumed to follow the order of the dates provided. It is assumed that any non-integer flight or simulation number in the data represents a flight cut short due to extenuating circumstances and can safely be omitted from the data set. While data available and funding are flexible resources, time is not. IP and student pilot time are the most important resource in the pilot training pipeline. Continuous student progress is assumed as the IP's primary responsibility. Therefore, sets of flight events organized by the IPs are always intended to advance student capability.

Conceptualization of AutoGradebook components, aside from the recommendation system, are out of scope for this research. Data provided excludes how individual grades were determined, when check rides occurred, how many times a specific event was attempted during an individual training exercise, weather patterns, and the time of day of each training exercise. Influence of any outside factors on event recommendation that are excluded from the data is also considered outside of scope.

1.5 Research Objectives

1. Evaluate current PTN data collection and storage practices to make recommendations for improvement.

2. Design a new metric to track student training progress more effectively than the current method.

3. Analyze and reveal patterns for sequential event set generation throughout a pilot training campaign.

4. Devise an algorithm that generates an appropriate set of flight events for an upcoming flight given student evaluation history.

5. Demonstrate objectives 2-4.

1.6 Research Contributions

An initial flight plan recommender system is defined, implemented, and tested. No other training flight plan recommender system appears to exist. Insights gathered from current operations motivated the proposal of a more effective data environment and student performance metric. The new student progress metrics are defined and employed to guide recommendation for student flight plans. Guidance for establishing an effective data environment provides a foundation for future data implementation.

II. Literature Review

2.1 Overview

This chapter discusses the origins and evolution of recommender systems, applications in research and industry, and common recommendation generation approaches. Additionally, this section discusses complications associated with implementing more personalized recommender systems.

2.2 A History of Recommender Systems

In any decision-making scenario, it is essential that all possible options are explored to make a good decision. Possible options for decision makers become more complex as a corresponding system grows in size, complexity, or influence. Historically, people have relied on peer and expert recommendations to simplify larger decision-making scenarios. As experts or decision makers seek more personalized recommendations, social methods of acquiring information cannot always provide specific enough advice. Computer-based recommendation systems introduce the ability to obtain more specified information or advice for a decision maker's interests [6].

As computer-based recommendation systems became standard practice in decision support, the use of automated recommender systems grew more common. Early automated recommender systems depended on hard-coded, user-provided specifications to filter through possible options and make suggestions. However, research has advanced the benefits of automatic recommendation systems for decision making processes to nearly a standard practice. Today, many online recommender systems do not require user input to generate recommendations. Instead, modern recommender systems often employ automatically recorded data from user activity to generate effective suggestions. [6]

A key component in improving consumer experience rests in effectively providing a simplified set of choices for a user. Therefore, effective recommender systems have become essential to the success of major E-commerce companies such as Amazon and Netflix [7]. Recommender systems are widely used in modern decision making scenarios and aim to capitalize on a variety of methods to provide effective recommendations to the targeted user.

2.3 Common Recommendation Generation Approaches

Baseline prediction methods must first be established in order to implement more personalized recommendation algorithms. Baseline methods include non-personalized methods such as data pre-processing and normalization [6]. Taking the average rating over all ratings in a system exemplifies a simple baseline method. Baseline predictors can also be enhanced for better results in ways such as combining mean values for a given item with average deviations from those values. The most common methods for generating user recommendations are collaborative filtering and content-based filtering.

Collaborate filtering methods rely on the assumption that users with highly correlated behavior would prefer similar recommendations. Similar users are grouped together to provide a reasonable prediction of active users preferences based on feedback ratings or user behavior within a system [6]. One of the original automated collaborative filtering algorithms is the k-nearest-neighbor(k-NN) collaborative filtering technique, which finds users with behavior similar to the current user and predicts ratings to the user based on similar users' preferences [8]. Two of the better performing methods of collaborative filtering include latent factor models and neighborhood models, which find relationships between users, items, or both, to highlight those important factors used to tailor recommendations [9].

Item-based collaborative filtering, another widely used method, relies on similarities between rating patterns of items rather than user behavior to make predictions.[6]. Companies that sell consumer goods, such as Amazon, often use item-based filtering to advertise goods that meet customer needs [10]. Netflix, and other companies looking to provide entertainment services tend to use hybrid approaches of item-based and k-NN methods to capture personalities of consumers rather than the functionality of a specific item [7].

Collaborative techniques are also used to guide population learning behavior. Particle swarm optimization (PSO), a heuristic global optimization method, compares the location of individual points within a population with the location of the best known point within the population. A point's fitness is defined as the overall best location it has achieved in the population space. The population's fittest point is the deemed population optima and returned as the estimated optimal point. The locations of individual points are adjusted according to individual inertia, the individual's fittest point, and the population's fittest point [11]. The algorithm redirects points back towards the population optima if they are adjusted past it. PSO is a proven approach for machine learning, classification tasks, neural network training, robot task learning, and other functions often accomplished using genetic algorithms [12].

Content-based filtering examines an individual user and produces future recommendations similar to items previously preferred by the user [6]. Information retrieval and information filtering are the primary tasks of any content-based filtering system. Vector spacing algorithms group items with similar feature information to create preference profiles for the user. Content-based user preference profiles are portrayed as a vectored combination of weighted item features. Multivariate techniques such as Bayesian classifiers, cluster analysis, decision trees, and artificial neural networks are examples of methods for user profiling [13]. Pandora, an online music streaming company, has had success applying a content-based filtering algorithm to recommend new music to users [7].

Deep learning models are composed of layers of artificial neural networks that look to exploit the unknown structures within data using multiple levels of connected weighted values. The ability for deep learning models to learn deep representations and abstractions from data has propelled deep learning model architectures to the leading edge of supervised and unsupervised learning tasks. The different type of neural network models that are suitable for different recommendation tasks can be looked at as neural building blocks for complex models. Deep neural networks can be composed of multiple neural building blocks that form one functioning model. The flexibility in modeling options introduces the ability to model vast amounts of complex data, providing an additional advantage for content-based recommendation tasks.

The PTN recommendation task must model the sequences of flight plans throughout training. The AutoGradebook recommender system must be able to model the temporal dynamics of a pilot training student throughout training in order to produce legitimate recommendations for flights maneuvers to be performed in upcoming flights. Extensive literature finds deep neural networks successful in a variety of sequence modeling tasks such as translation, natural language processing, music generation, dialogue generation for chatbots, weather prediction, next-item/basket prediction and more [14]. Two deep learning model structures that have proven successful in capturing temporal data trends for prediction purposes are the Long Short-Term Memory Recurrent Neural Network (LSTM RNN) and Temporal Convolution Network (TCN) [15].

Recommender systems seek to maximize the expected utility of recommendations as a whole rather than of the individual items. However, using a single recommendation method may not generate effective predictions in all scenarios. Different models perform better given different scenarios and data available. Various individual models may capture a unique part of what a true prediction or recommendation should be. However, one model rarely captures the entire truth. When building generalizable models, diversity is strength. It is possible to combine modeling achitectures to formulate a hybrid approach to recommendation [6]. Hybrid filtering is a specific application of model ensembling where the predictions of a set of different models, or neural building blocks, are combined to produce better predictions. In hybrid filtering, methods are performed independently and combined using weights or preliminary cascading techniques [13]. In 2011, the Recommender Systems Handbook [16] was published, providing an in depth overview of recommendation systems and methods.

2.4 Potential Complications of Personalized Recommendations

Recommender Systems are not a one-size-fits-all solution to making decisions, so they must be personalized to individual systems or applications. This individualization often results in some complications. Providing more personalized recommendation based on user interaction can introduce inconsistencies in recommender system generalization ability. In a pilot training environment, inconsistencies in student activity or performance can introduce overall inefficiencies in recommender performances if the student does not show any progression trends.

There are two primary types of users in any recommender system. From an advertising standpoint, the ideal users are known as white sheep. White sheep are users that perform similarly to many other users. Users that are less predictable, known as black sheep, show a low correlation in behavior compared with almost all users. Recommender systems generally perform poorly on black sheep users and their low correlation of behavior introduced into training data causes diminished recommendation quality for other users in the system [17]. In pilot training, students who outperform their peers at the beginning of training, but underperform later on, or vice versa, may act as black sheep.

Another common complication in recommender systems is the introduction of new data. As a new program, PTN continuously records new training data. Many recommender systems discussed in the literature assume that the nature of the database being used to train a recommender system algorithm is static. One proposed method to dealing with dynamic databases it to implement a two-stage approach to identify relevant recommendation options and then provide user specific recommendations based on preferences [18]. K-NN collaborative filtering methods suffer from scaling complications when introducing new data while item-based filtering methods do not [6]. Recommender systems often have very specific function, so complications arise if they are used in more general environments or are applied to a broader set of problems. Despite these complications, recommender systems are used widely in industry.

Model performance evaluation can also be difficult when building recommender systems. Mathematical evaluation methods tend not to correctly measure the performance of any given recommender system. Metrics such as root mean square error fail to recognize the practical use of a recommender because they do not measure the impact that recommender systems have on the user. Evaluation of recommender systems typically focuses on a predicting task and a recommendation task. An item is defined as all possible options a recommender provides to the user. The predicting task regards an algorithm's ability to identify the value of an item to the user. The recommendation task involves an algorithm's ability to produce the best possible list of items according to a user's needs. If a system is being used as a decision support tool, the user may want the system to actually make suggestions or select the best decision given the current environment. Instead, it may be more relevant to evaluate a recommender system based on metrics such as serendipity or diversity in recommendations. Serendipity can be measured as recommendations that a user likes but did not think of initially. Diversity in recommendations assures that the user gets exposed to all possible items [6].

III. Data Details

3.1 Overview

This chapter describes the data provided by the PTN program and initial data cleaning needed for analysis. Recommendations for future data collection and data storage are offered.

3.2 Data Description

PTN leadership provided raw data collected from their first pilot training class. The raw data set consists of the scores received on every graded event performed during each training exercise for the nineteen students in the original PTN class. The data includes information on 128 individual flight events possible of being executed during pilot training. Training events are distributed into 10 different event categories: basic, patterns, contact, instruments, basic formations, tactical formations, low-level, four-ship formations, combat air forces (CAF) introduction, and mobility air forces (MAF) introduction. Only a subset of all possible flight event can be performed during each training exercise. Students are identified using an identification (ID) number within the data set. Each record in the data represents the information for a single training exercise for a given student. The data consists of the student ID number, training exercise date, information on all 128 possible flight events, and the device used for training (simulation or flight). Additionally, each event falls into one of the aforementioned ten categories defined by PTN. The ten categories and their corresponding events are listed in Table 1.

Table 1. List of graded flight events by category provided by PTN

Event Categories	Graded Flight Events
Basic	Mission Analysis/Products, Ground Ops, Takeoff, Departure, Basic Aircraft Control, Cross-Check, Enroute Descent/Recovery, Inflight Checks, Inflight Planning, Clearing/Visual Lookout, Communication, Risk Mgmt/ Decision Making, Situational Awareness, Task Management, Emergency Procedures, General Knowledge
Patterns	Overhead/Closed Pattern, Visual St-In, Landing, No-Flap Landing, Go-Around, Emergency Landing Pattern
Contact	G-Awareness, TP Stalls, Slow Flight, Power On Stalls, Contact Recoveries, Spin Recovery, Aileron Roll, Barrel Roll, Pitchback / Sliceback, Cloverleaf, Cuban Eight, Immelmann, Lazy Eight, Loop, Split S
Instrument	Vertical S, Unusual Attitudes, Steep Turns, Intercept/Maintain Arc, Fix to Fix, Holding, Full Procedure Approach, Non-Precision Final, Precision Final,Circling Approach, Missed Approach, Night Landing
Basic Formation	Wing Takeoff, Interval Takeoff, Instrument Trail, G-Warmup/Awareness, Lead Platform, Pitchout(Both), Fingertip(Wing), Route(Wing), Fighting Wing(Wing), Straight Ahead Rejoin, Turning Rejoin, Overshoot, Echelon(Wing), Breakout(Wing), Lost Wingman(Both), Extended Trail(Wing), Position Change, Formation Approach(Both), Formation Landing(Both), Battle Damage Check,Flt Integrity/Wingman Consideration
Tactical Formation	Delay 90, Delay 45, Hook Turn, Shackle, Cross Turn, Fluid Turn, Tactical Rejoins, Fluid Maneuvering, Tac Initial
Low-Level	Course Mx, Course Entry, Time Control, Altitude Control, Checkpoint ID, LL GPS Integration, Tactical Maneuvering LL Lead Change
4 Ship Formation	Four Ship Admin, Fluid 4, Box Formation, Offset Box, Wall, 4-Ship Fingertip, Straight Ahead Rejoin, Turning Rejoin
CAF Introduction	Heat to Guns Setup, Heat to Guns Maneuvering, Fuel Awareness/Management, Advanced Handling, Perch Setups, Maneuver Selection, Offensive Fighter Mnvr Exec, Defensive Fighter Mnvr Exec, CZ Recognition, Air to Air Weapons Employ, HA Lead Turn Exercise, HA Butterfly Setups, HA BFM Flt Analysis, SA Conventional Range, SA Tactical Range Proc, SA Safe-Excape Maneuver, SA Threat Reaction, SA Weapons Employment, Air to Ground Error Analysis, TACS/JFIRE Procedures, Air to Ground 2-Ship Mutual Supt
MAF Introduction	Mission Management, VFR Arrival, Tanker Procedures, Reciever Procedures, Airdrop Procedures, Crew Coordination, Single Engine Approach, Single Engine GA/Missed Appch, A/R Overrun, A/R Breakaway, FD/AP Operations, FMS Operations

Each student's training exercises are ordered by date and labeled to create a training exercise data point for each record. Unrecorded factors, such as weather or resource availability, introduce possible volatility in specific training dates for students. Using training exercise numbers rather than a calendar date to track progress introduces a standard temporal metric between students. The training exercise metric represents the total number of training exercises that a student has completed in training up to that associated flight record.

While the training exercise calendar dates are included in the data, specific time stamps of when training exercises started, when specific graded events were performed throughout a training exercise, and timestamps for when training exercises concluded are not provided. Introducing more temporal data may provide further insights to how events should be paired within a single training exercise as well as the order in which events should be performed. The absence of this additional temporal data inhibits the ability to fully make data-driven decisions regarding how graded exercises should be paired within a single training exercise. For example, contact events are not typically performed at night, but a training exercise may start before sunset, where a student may be scored on contact events, and continue into the night so a student may be scored on a night landing. There is currently no data collected connecting exercises or events to the time of day that they are performed. There is also no data on the order of events performed within each training exercise. Therefore, it is assumed that there are no event pairing constraints for the given set of 128 possible graded flight events.

3.3 Preliminary Data Preparation and Collection Recommendations

Preliminary data cleaning was performed by AFWERX. However, further data cleaning was necessary to prepare the data for exploratory data analysis. Ambiguous and inconsistent data labeling, improper data formatting, inconsistent recording, and incomplete recording were among the most common issues found within the data.

Ambiguous data labels were converted to more specific labels. An example of this correction includes replacing the data category label “Training Day”with the label “Training Date”. The term “Training Day”does not specify whether the data being provided is the ordered day since the start of training, the day that the data was recorded, or the calendar day that the flight with corresponding scores took place. Ambiguity is avoided by providing more specific labeling, such as “Training Date” in the provided example. Specific data documentation was unavailable from PTN and AFWERX, making ambiguous labeling more challenging to decipher. Future data collection should include more context specific labeling and an associated data dictionary to increase functionality and employability of all recorded data.

All data provided were collected from a manual input system; IPs manually inserted each individual data record. This approach structures all data as text, regardless if the proper data format should be numeric, non-numeric, or temporal. Fully manual input can also result in inconsistent input. Inconsistent spelling and additional unnecessary characters were often found in the data, creating confusion. These discrepancies were fixed manually in preliminary data cleaning. A “point and click” graphical user interface would provide a simple solution to this problem and introducing more convenience and efficiency to the data collection process in the future.

The data also contained various incomplete or duplicate records. For most incomplete records, there were associated complete records occurring on the same date. This suggests that software defects, network issues, or recording errors may cause incomplete and duplicate records when IPs reinsert recorded training results. Some records that included less than five evaluation scores appear to represent evaluation updates for specific graded events rather than completed training exercise records that the current data structure supports. Inclusion of duplicate or incomplete records of training exercise data may result in skewed, and less effective, recommendations. Ambiguity in record completeness exists given that there is not a definitive number of graded events that may occur each flight. All incomplete or duplicate records were removed from the dataset to limit biased recommendations that do not lead to student graduation. Network issues and software defects are often unpredictable, but recording a data point indicating a complete training exercise record would distinguish between complete and incomplete records and provide insight into the credibility of future records. Incorporating a confirmation option before officially submitting a record provides an automatic solution to the addition and recording of the suggested completed record data point. In the occurrence of a mishap, automatically saving recorded progress and preventing the IP from creating a new record until the incomplete record is confirmed or canceled may provide a solution to the recording of incomplete or duplicate data.

Upon initial cleaning, additional data were recorded in order to more effectively capture temporal trends of student progress throughout training. Outside factors such as days off, unsuitable weather conditions for flight or performing specific events, limited aircraft inventory, and limited IP availability contribute to unpredictability of when the next evaluated training exercise will occur. The training dates of individual flights are vulnerable to unknown or unrecorded factors. Therefore, a more generalized recording of flight order provides a more appropriate approach to track temporal trends in student progress. A “Training Exercise” data category was calculated to represent the total number of evaluated training exercises that a student has performed leading up to the recorded exercise. Recording temporal data by training exercise provides enough information to show student progression through training based solely on when and how skillfully events were executed over time, independent of unknown influential factors.

Currently, overall student progress in flight training is monitored using a cumulative MIF super score. The cumulative MIF score is calculated by taking the summation of the maximum recorded scores of each individual training event introduced up to the current training exercise. The cumulative MIF metric is also used to compare performance between students throughout pilot training. Tracking a student's progression through pilot training must account for the student's depth and breadth of event knowledge. A student's event depth is defined as how skillful the student is in a particular event. Depth is measured by a student's highest received score in a graded event. Event breadth is defined as how many of the 128 total possible events the student has been introduced to throughout training. Breadth is measured using the total number of individual events and the number of event categories fully introduced to the student. The current use of the cumulative MIF score does not distinctly account for both the breadth and depth of student progression. Therefore, using the cumulative MIF metric leads to ambiguity in true student performance.

The following is a simplified example emphasizing the ambiguity of the current cumulative MIF metric. Suppose an IP would like to compare progress between two pilot training students, Student A and Student B. Both students have performed an equal number of training exercises. Student A has been introduced to, and received, a score of excellent on only two individual events. Student B has been introduced to, and received, a score of unsatisfactory on eight individual events. Both students in this scenario have a cumulative MIF score value of 8 points. There is no way for the IP to adequately distinguish the progress of the two students from the cumulative MIF metric alone. However, further examination of each student's grade sheet shows that Student A has a proficient grasp on both of the exercises they have been introduced to whereas Student B is consistently performing at unsatisfactory levels in all aspects of training.

For this work, the cumulative MIF metric was replaced by a new performance metric, the Forward Progress Score (FPS), that distinctly incorporates both training depth and breadth. The FPS greatly enhanced the capability of the recommender algorithm conceptualized, prototyped, and tested in this research. FPS calculation and rational are further defined in the Modeling Approach Chapter.

3.4 Summary of Data Recommendations

The data cleaning process yielded various recommendations for developing a data environment better designed for future data implementation. Those recommendations are:

1. Improve the temporal aspect of data collected.
2. Timestamp training exercises.
3. Implement stronger data formatting.
4. Implement standard data labels.
5. Improved interface for data entry.
6. Incorporate record confirmation components.
7. Add outside data to support the training data collected.
8. Consider the adoption of Forward Progress Score metric.

IV. Exploratory Data Analysis

4.1 Overview

IPs must take into account various student and environmental factors when creating training flight plans. This section dives into the data to uncover graded event frequency and evaluation trends throughout the PTN program. Trends revealed in this section establish the foundation for the modeling approach taken in this research.

4.2 Event Frequency Over Time

Despite sparse data, some general trends for event occurrence throughout training are present. Event frequencies suggest that event category occurrence and proficiency are not bound to a specific timeline in PTN. Frequency charts, shown in Appendix A suggest that event recommendations follow patterns regarding the training exercises when they are performed. An example of the frequency chart for a single event is shown in Figure 1.

The data does not present a specific order in which events should be introduced, but events within categories tend to follow similar patterns of what training days they occur.

No pairing restrictions have been established based on the flexibility of the PTN training program and training exercise structuring. Event frequencies alone do not provide enough evidence to establish specific events pairings regardless if multiple event categories may peak at similar training exercises. However, analysis of individual records suggest some event general category pairings may exist. All observations of event pairings are generalized observations from the data and not hard set constraints for recommendations.

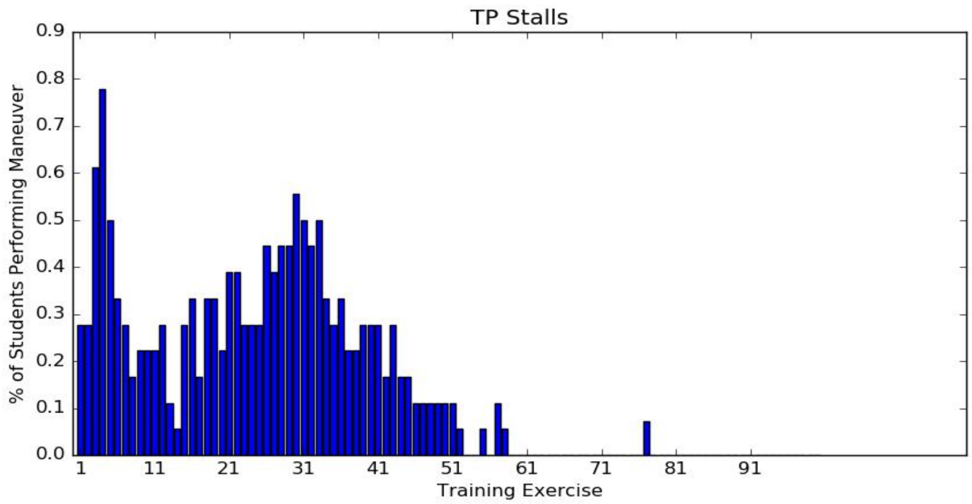


Figure 1. Frequency of Occurrence Example for the TP Stalls Event

Basic events are the most commonly performed event, independent of proficiency. Basic events are performed on almost every training exercise for all students.

Instrument events do not show continuous patterns. Vertical S, unusual attitudes, and steep turns tend to occur in early training exercises, and are revisited later if needed. The rest of the instrument events are performed continuously throughout training, with peak occurrences between twenty and thirty training days. An emphasis in instrument training also peaked during training occurring after 90 training exercises. Late peaks in occurrence suggest that the IP emphasized specific events because those events are essential to program graduation and the IP did not feel that skill levels in those events were yet up to expectations.

Contact event frequency peaks around training exercise 10 and training exercise 30. They seem to occur less during the end of training. Some contact events and instrument events are performed in the same training exercise, but there never seems to be a shared focus of these two categories. If events from both categories were performed in a single training exercise, there was always a heavier inclusion of one category over the other.

Table 2. **Grouped Basic Formation Events Based on Event Occurrence Frequencies**

Group A	Group B
Wing Takeoff	Interval Takeoff
G-Warmup / Awareness	Fighting Wing (Wing)
Lead Platform	Instrument Trail
Pitchout (Both)	Turning Rejoin
Fingertip (Wing)	
Route (Wing)	
Straight Ahead Rejoin	
Overshoot	
Echelon	
Breakout (Wing)	
Lost Wingman (Both)	
Extended Trail (Wing)	
Position Change	
Formation Approach (Both)	
Formation Landing (Both)	
Battle Damage Check	
Flt Integrity / Wingman Consideration	

Basic formation events split into two distinct groups based on temporal trends, as shown in Table 2. Events within group A tend to be introduced earlier in training, around training exercise 10, and are then revisited later in training. Events in group B tend to only be performed after training exercise 40. Events in the tactical formation, low-level, and four-ship formation categories, with the exception of the straight ahead rejoin event, tend to occur in later training exercises, as well. Training exercises evaluating low-level events tend to also evaluate tactical formation events as opposed to basic formation events.

Pattern events show no temporal trends and are incorporated with events from all other categories. Pattern events and low-level events are rarely evaluated in the same training exercise.

CAF and MAF events are all performed towards the end of training. These events are included to provide students with an introductory insight to the next phase of pilot training. All MAF and CAF events were performed after training exercise 60.

Students that graduated with fewer total number of training days did not receive as much exposure to MAF and CAF events as those that took longer to graduate, indicating that individual student performance may not be the only driving factor determining when a student graduates. For example, all students completing training in under 70 training exercises were not introduced to any MAF or CAF events at any point in their training. Student performance in four events from the MAF category were not evaluated in any recorded training exercises from the first PTN graduating class. The group of four non-evaluated events includes single engine approach, single engine GA/missed approach, A/R overrun, and A/R breakaway.

Inconsistencies in expectation for which events are introduced between students' campaigns creates even more ambiguity to the necessary skill requirements and expectations for students to advance to the next phase of pilot training. Ambiguous graduation expectations may be leading to inconsistent performance of graduates. Personalization of progress and standardization of expectations must be balanced in order to create a consistent production of capable pilots and an effective pattern of flight event recommendation.

4.3 Temporal Progress Trends

4.3.1 Variation in Training Lengths

The data indicates that there are significant variations in the necessary amount of training exercises individual students may need to graduate the PTN program. Six students completed training between 60 and 79 training exercises while twelve students completed training between 80 and 100 training exercises. The minimum number of training exercises required for a student to graduate in the first PTN class was 60. The maximum number of training exercises required for a student to graduate in the first PTN class was 100. The average and median number of training exercises required to graduate in the first PTN class were 83 training exercises and 87 training exercises, respectively. Individual training lengths per student are shown in Table 3. Analyzing length until completion excluding MAF and CAF events may provide valuable insight on the true length of training. However, there is no way to identify training completion solely based on the data other than a halt in training data for each student. A strict, standardized completion requirement must be established before training completion length can be appropriately projected.

Table 3. Training Exercises per Student

Student ID	Training Exercises to Graduate
11	100
18	95
12	94
15	94
1	93
10	92
13	91
5	90
9	87
3	86
14	80
7	80
4	79
2	77
8	69
19	69
17	66
16	60

The data in Figure 2 shows a left-skewed density distribution for overall training length. A left skewed density distribution suggests that students are less likely to finish training early. Factors influencing total graduation time remain unknown because of inconsistencies in event exposure. Figure 2 shows a histogram representing the percentage of students graduating on specific training days and a corresponding density curve to estimate what the probability of finishing training on a specific training

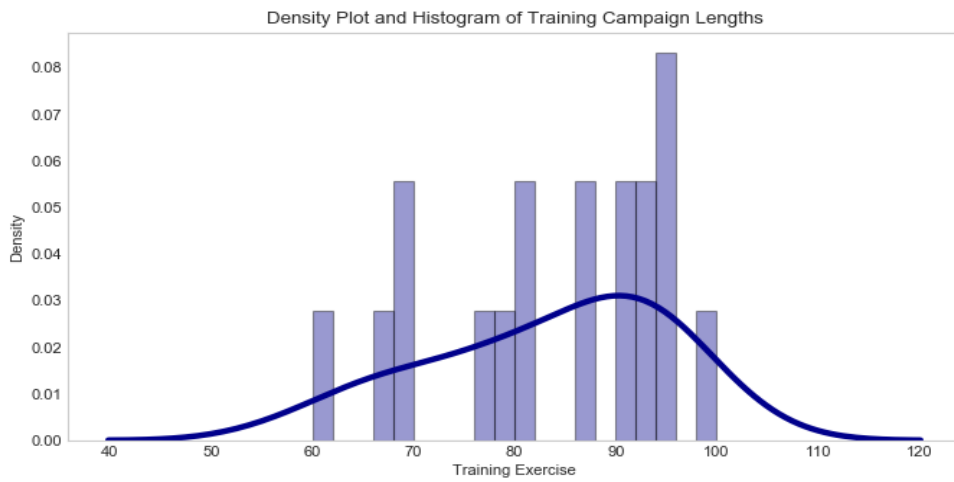


Figure 2. Density plot showing the probability of finishing training on a given number of training exercises.

exercise may look like given a larger dataset. Variance in training length supports PTN claims to make pilot training more personalized to individual student training progress and performance. Trends in overall student progress and individual event performances were broken down to better understand the event recommendation process performed by IPs before each training exercise.

4.3.2 Student Evaluation Performance Over Time

Maximum, average, and minimum score statistics for each training exercise were calculated for all events. Scoring statistics for each training exercise were calculated using the scores of all students that performed equal to or greater than the specified number of training events. Figure 3 shows an example of the change in student evaluation statistics over time for a single flight event.

Volatility in score statistics occurs because not every event is performed on every training exercise. Students who reach proficiency in an event tend not to be reevaluated on that same event during their next flight. The students performing that

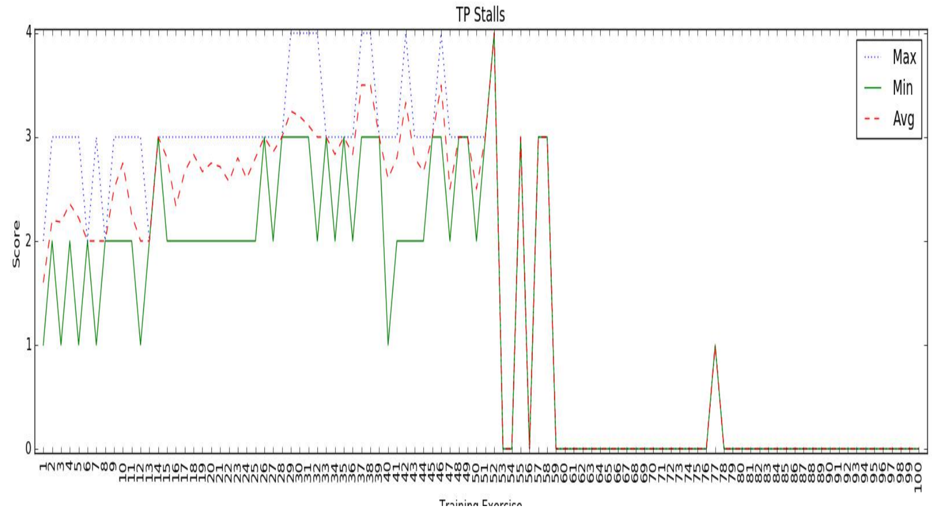


Figure 3. Evaluation Statistics Example for TP Stalls Event

event in the next flight are generally students that have not reached proficiency yet. Limited data samples also contribute to variability. Performance evaluations in the TP Stalls event continue to improve up until training exercise 53, where the average score is equal to the maximum score for the event. After training exercise 53, the scores go down to 0, indicating that no students performed the associated event on the associated training exercise. It was found that fewer students performed an event when all three statistics in Figure 3 showed equal values on a given training exercise. This occurred more often in later training exercises as more students got close to graduating. Often, if the minimum, average, and maximum were equal, only a single student performed the flight event on the corresponding training exercise. Visualization of temporal evaluation statistics for all events can be found in Figures 14-141 in Appendix A.

Trends for reaching proficiency, or individual MIF, were examined for each possible graded event. Table 4 shows the minimum, median, average, and maximum number of training events required to establish proficiency in each graded event. The values are representative of the students in the original PTN class that reach proficiency. Any students that never reached proficiency in a specific event were not included in the descriptive statistics for reaching proficiency in that event. Table 4 provides insights regarding the number of training exercises expected for students in a new PTN class to reach proficiency. Missing values indicate that no student reached proficiency in the associated flight event during their training campaign.

4.3.3 Overall Student Progress

Table 5. Point Allocation Per Event Graded Evaluation

Recorded Grade	Definition	Point Equivalent
E	Excellent	4
G	Good	3
F	Fair	2
U	Unsatisfactory	1
NG	No grade	0
N/A	No recorded data	0

The MIF super score was calculated for every training record per student to analyze how well MIF can track training progress. Points were allotted according to how well a student scored on every completed event during a flight. At each training exercise, a student's individual event MIF scores were calculated according to maximum grades and the corresponding point allocations in Table 5.

A score of 0 was recorded for all events not performed or any events that were performed but not graded during a training exercise. A visual representation of all student MIF super scores over time is shown in Figure 4.

Each line in Figure 4 represents the cumulative performance throughout training of one of the 18 students that graduated from the first PTN class. There was only one student who failed out of the first class of PTN. The disenrollment was assumed based on an absence of recordings after training exercise 26 and poor evaluation performance leading up to that point. Data from that student were omitted from the dataset used

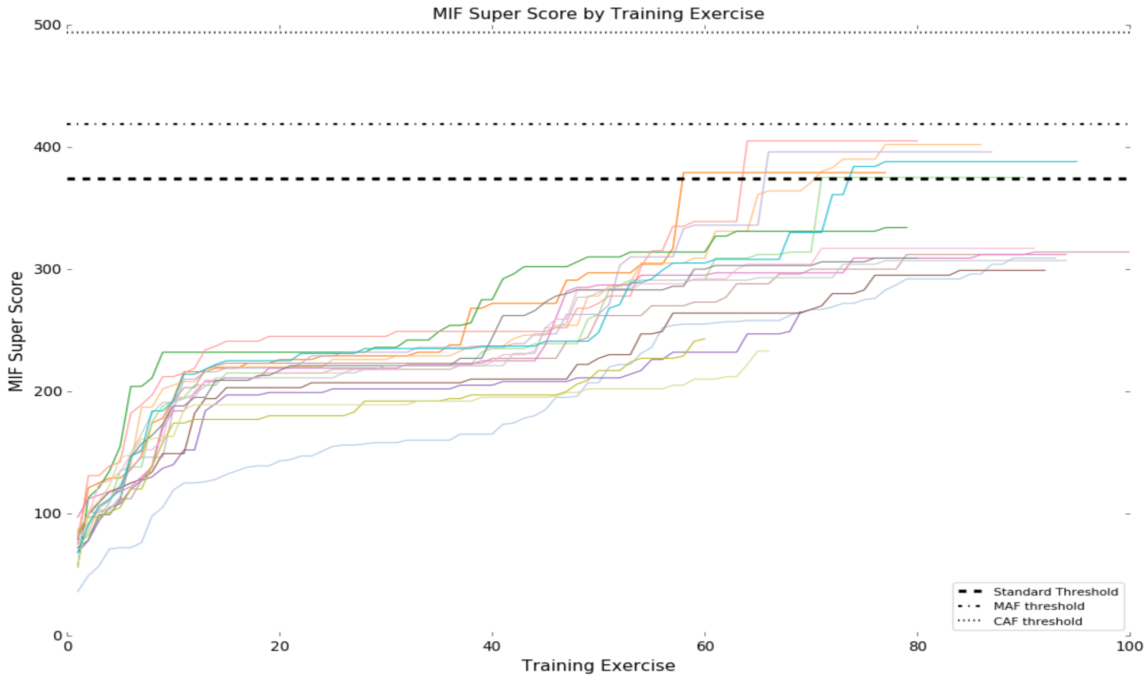


Figure 4. Cumulative student performance over time cumulative MIF metric for first PTN class including CAF and MAF events

in model development because the training path does not represent one that can be used to graduate pilot training students.

The standard MIF threshold, represented by the thick dashed line, is the maximum cumulative MIF score that a student can accrue by excluding MAF or CAF events from the evaluation. The MAF threshold is the maximum cumulative MIF score a student can accrue excluding CAF events from the grade sheet. The CAF threshold is the maximum cumulative MIF score a student can accrue when including all possible events in their training campaign.

Any students who finished training above the MAF threshold were on the CAF track. Any students who finished training above the CAF threshold were on the CAF track, but also performed some MAF events. Figure 4 shows student progress, but there is nothing accounting for depth and breadth of student progress, so the comparison between students is difficult. For example, there is no distinguishing between students that have performed poorly on many events and students that have performed well on a few events. Although students can receive a score of excellent and receive 4 points on every event, some events only have a maximum MIF score of 3. This indicates that it is possible for students to score above the proficiency threshold. The ability to go above the max individual MIF in events also increases progress ambiguity because a score resting exactly on the performance threshold does not necessarily mean that the corresponding student is proficient in every event.

PTN leadership and the AFWERX-Austin team have hypothesized that the sharp increases followed by extended periods of slow increase support a notion that students distinguish themselves when new events are introduced. According to this hypothesis, Figure 4 shows that many new events were probably introduced to the students before training exercise 15 and around training exercise 50. The performance of students that grasp events quickly has a more positive rate of change when introduced to new events.

Each student was found to progress towards the proficiency of individual events at a personalized pace. Extreme training personalization leaves each set of event recommendations up to the subjectivity of the IP, diffusing category based progress trends that may occur in traditional undergraduate pilot training. However, general progress trends in depth present themselves. Stagnating progress suggest that there are alternative factors influencing student progress that are not represented by the current progress metric. Also, Figure 4 suggests that the value of improving scores in event skill follow a nonlinear relationship because it is more difficult to advance from good scores to excellent scores than to improve from unsatisfactory scores to fair scores.

Figure 4 shows how the length of student training can vary tremendously. The PTN program timeline varies from student to student. Thus, a model making recommendations cannot rely on a set curriculum timeline. Instead, a model must be personalized to individual student's pace of progress over time.

In the current state of the PTN program, not all students are required to be introduced to MAF and CAF events to graduate. Therefore, all data from the MAF and CAF tracks within training were omitted to provide a standardized visual of events introduced to each student. Figure 5 displays the cumulative performance over time for each student disregarding CAF and MAF data.

Figure 5 suggests that every student does not breach the standard threshold. Failure to meet MIF proficiency standards may be the result of failure to accurately record data or subjectivity in the MIF proficiency measure introduced by the human component of the evaluation process. Even though there is an expected MIF score to establish proficiency for every event, every student that graduates does not reach that score. One possible explanation for this is that IPs have the power to waive a students sub-MIF performance if they think the student's performance is satisfactory for graduation despite recorded grades.

The Forward Progress Score (FPS) was designed to better model student progress by incorporating more depth and breadth aspects of training into a single metric. Achieving proficiency in each event is assumed as the primary goal for each student. The MIF super score represents student grades, but fails to clearly describe student progress towards proficiency. The FPS uses a percent value of the individual max MIF scores to establish a variable representing student progress toward proficiency in each event. Applying percentages of total progress towards a set goal addresses skill depth in the campaign toward overall proficiency more appropriately than simply considering recorded grades.

Visuals representing student FPS score over time, including and excluding MAF and CAF event evaluations, are represented in Figure 6 and Figure 7, respectively.

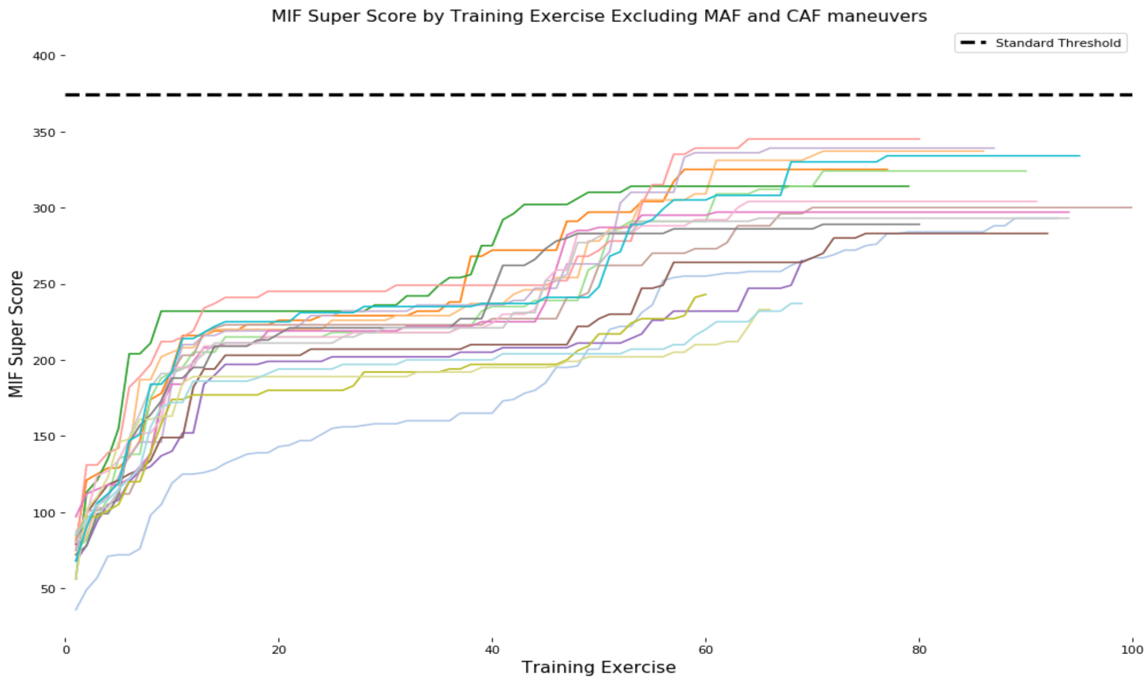


Figure 5. Cumulative student performance over time using cumulative MIF metric for first PTN class excluding CAF and MAF tracks

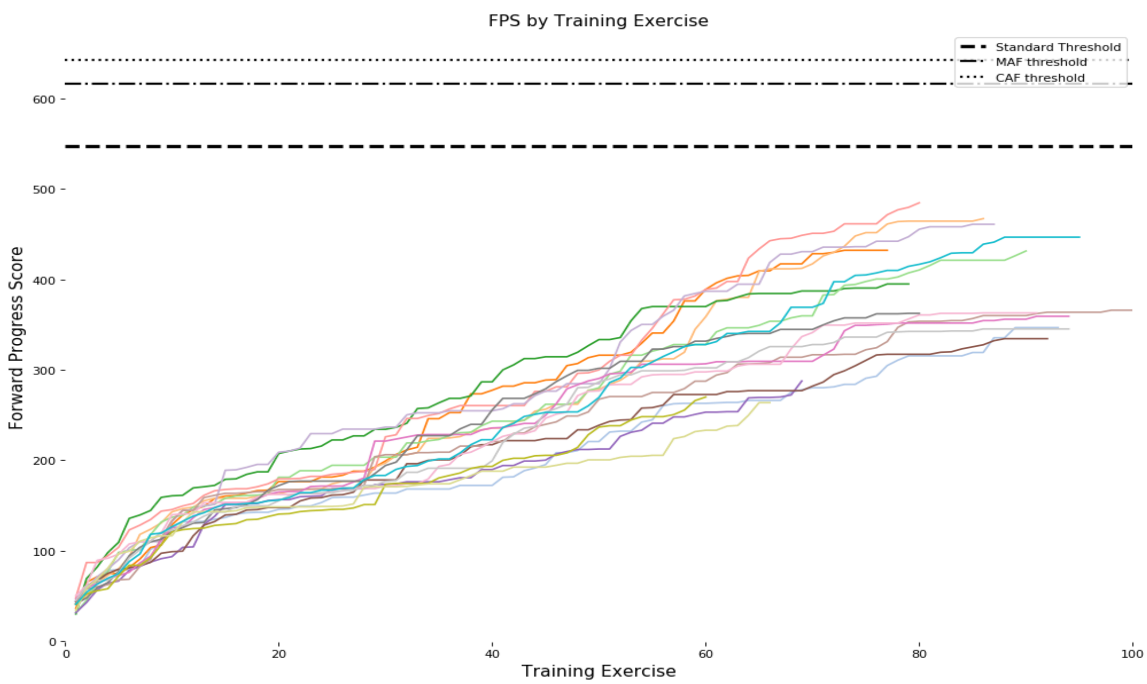


Figure 6. Cumulative student performance over time using FPS metric for first PTN class including CAF and MAF events

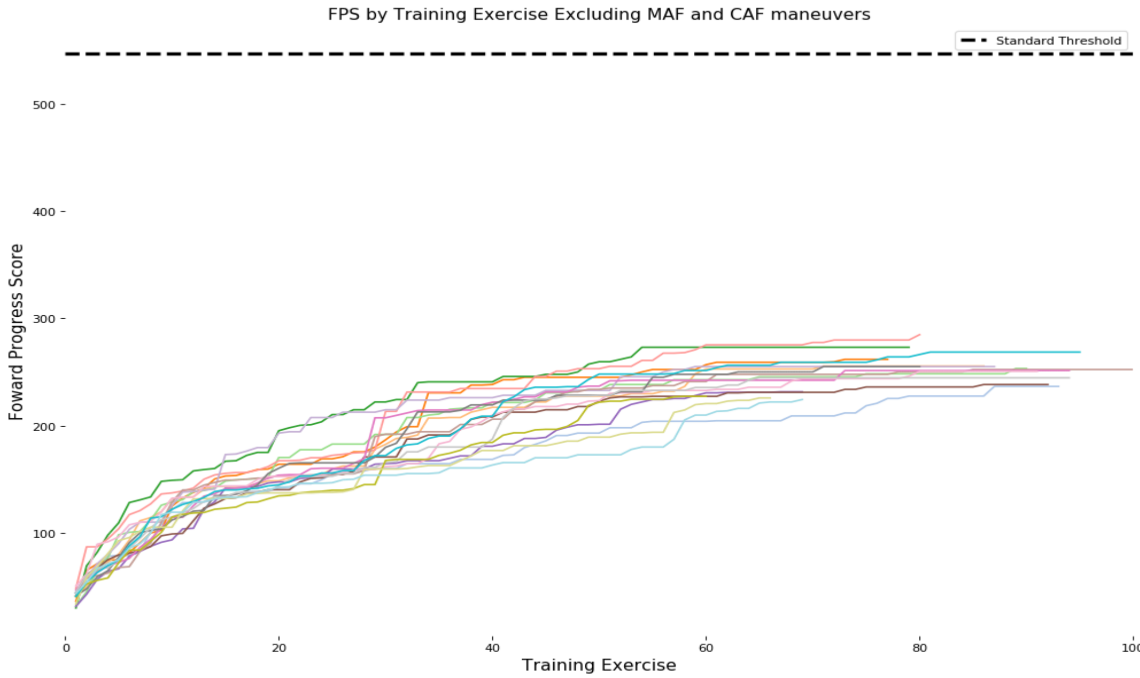


Figure 7. Cumulative student performance over time using FPS metric for first PTN class excluding CAF and MAF tracks.

Unlike Figure 4 and Figure 5, Figure 6 and Figure 7 show a consistent progression of student performance throughout training. Visual results indicate that the performance stagnation seen in Figure 4 and Figure 5 may be a result of the MIF metric's inability to account for all components of student progress. The components of student progress are explained in detail in the Modeling Approaches Section. Furthermore, the comparison between the figures tracking MIF score and the figures tracking FPS score shows that FPS better depicts continuous student progress throughout the training campaign.

4.4 IP Recommendation Trends

Further analysis was performed examining the order in which IPs introduced event. Averages for ten variables corresponding to training event selection were examined and portrayed in Tables 6 and 7. Table 6 shows how many events are performed on average at each training event and how new maneuvers are introduced as more training exercises are performed. # Total Events Used represents the number of 128 possible flight events that have been evaluated on the corresponding training exercise. # Current Events Used represents the number of events evaluated in an exercise. # New Events Introduced represents the number of events that were evaluated for the first time. # Used Events performed corresponds to the number of events that have previously been evaluated and are being evaluated again. Columns in Table 7 indicate statistics on how many events are taken away and added in the short term. For example, New from Past 1 and Deleted from Past 1 represent the number of events evaluated in the exercise that were not evaluated in the previous exercise and the number of events not evaluated in the exercise that had been evaluated in the previous exercise.

The data shows that IPs evaluate students using between 20 and 30 events on each training exercise. Out of those events, about a quarter are events that did not occur in the previous exercise. On average, a few events appearing in any exercise had not been evaluated within the previous 3 exercises. Table 7 also shows that the number of new events from the previous exercise and the number of events removed from the previous exercise are relatively close on average, suggesting that IPs might be performing 1-for-1 event swaps when creating new flight plans for each training exercise. Events are introduced more frequently in the beginning of a training campaign, but become more sporadic throughout a campaign.

Ordered lists containing the training exercises when events were introduced for each student were produced. Analysis of event introduction orders did not provide any insights for a standardized order of introduction between events. The personalized nature of PTN supports the lack of standardized event introduction.

The progress curves in Figure 4 suggest that the advancement in individual event performance follows a nonlinear path. An in-depth analysis performed on individual students uncovers that students tend to spend the least amount of time, in training exercises, at the unsatisfactory level. Often, students may never receive an unsatisfactory score on an individual event. Students spent the most time at the Good level for each graded event. This observation is consistent throughout the data. IP bias implemented to reassure that a student is truly proficient in a graded event before officially recording their skill level is the hypothesized cause of this issue.

Table 4. Statistics for reaching proficiency across original PTN class measured by training exercise

Graded Event	Min	Median	Mean	Max	Graded Event	Min	Median	Mean	Max
Mission Analysis/Products	13	31.5	33	56	Extended Trail (Wing)	56	66	66	77
Ground Ops	14	28	32	64	Position Change	28	51	49	68
Takeoff	1	28.5	28	56	Formation Approach (Both)	54	57.5	58	61
Departure	3	28	29	57	Formation Landing (Both)	49	49	49	49
Basic Aircraft Control	13	36	38	79	Battle Damage Check	45	55	57	72
Cross-Check	19	33	39	86	Flt Integrity / Wingman Consideration	31	47.5	49	73
Enroute Descent / Recovery	14	31	33	74	Delay 90	41	52	56	77
Inflight Checks	14	31	32	57	Delay 45	52	61	63	77
Inflight Planning	19	34.5	38	74	Hook Turn	52	57	61	77
Clearing / Visual Lookout	14	30	33	74	Shackle	49	61.5	62	77
Communication	13	29	32	67	Cross Turn	54	59	61	75
Risk Mgmt / Decision Making	12	29	30	53	Fluid Turn	77	77	77	77
Situational Awareness	14	32.5	36	86	Tactical Rejoins	44	61	56	62
Task Management	19	31.5	35	76	Fluid Maneuvering	54	61	61	68
Emergency Procedures	19	31	38	86	Tac Initial	47	53.5	56	69
General Knowledge	19	29	37	79	Course Mx	47	66	65	88
Overhead/Closed Pattern	14	30	30	52	Course Entry	38	55	60	88
Visual St-In	7	32	26	33	Time Control	41	51	54	70
Landing	2	28	27	45	Altitude Control	50	52	60	88
No-Flap Landing	48	58	64	88	Checkpoint ID	51	63	64	88
Go-Around	19	34	36	60	LL GPS Integration	49	58	62	88
Emergency Landing Pattern	11	17	18	31	Tactical Maneuvering	51	66	63	76
G-Awareness	28	32.5	36	57	LL Lead Change	52	67	63	76
TP Stalls	28	30.5	33	45	Four Ship Admin	62	62	62	62
Slow Flight	-	-	-	-	Fluid 4	56	56	56	56
Power On Stalls	22	33	33	45	Box Formation	-	-	-	-
Contact Recoveries	21	30	33	57	Offset Box	62	62	62	62
Spin Recovery	2	7.5	10	36	Wall	70	70	70	70
Aileron Roll	16	28	28	40	4-Ship Fingertip	-	-	-	-
Barrel Roll	29	39.5	40	53	4-Ship Straight Ahead Rejoin	-	-	-	-
Pitchback / Sliceback	28	28	28	28	4-Ship Turning Rejoin	62	62	62	62
Cloverleaf	13	29.5	30	42	Heat to Guns Setup	65	69	69	73
Cuban Eight	3	32	28	38	Heat to Guns Maneuvering	59	65	66	72
Immelmann	11	32	32	43	Fuel Awareness/Management	59	69	70	81
Lazy Eight	28	40	40	56	Advanced Handling	-	-	-	-
Loop	27	31	32	40	Perch Setups	59	65	66	73
Split S	4	37	36	57	Maneuver Selection	61	65	67	73
Vertical S	-	-	-	-	Offensive Fighter Mnvr Exec	62	66	67	73
Unusual Attitudes	16	19	19	22	Defensive Fighter Mnvr Exec	64	73	73	81
Steep Turns	71	75	75	79	CZ Recognition	65	70	71	78
Intercept / Maintain Arc	31	53	54	84	Air to Air Weapons Employ	-	-	-	-
Fix to Fix	4	19	22	61	HA Lead Turn Exercise	-	-	-	-
Holding	29	48.5	46	62	HA Butterfly Setups	-	-	-	-
Full Procedure Approach	27	37	38	56	HA BFM Flt Analysis	72	74.5	75	79
Non-Precision Final	26	31	36	78	SA Conventional Range	-	-	-	-
Precision Final	1	29.5	30	55	SA Tactical Range Proc	-	-	-	-
Circling Approach	16	58.5	55	86	SA Safe-Excape Maneuver	-	-	-	-
Missed Approach	11	45	45	63	SA Threat Reaction	77	77	80	87
Night Landing	17	37	34	48	SA Weapons Employment	76	85	83	89
Wing Takeoff	14	43	46	75	Air to Ground Error Analysis	-	-	-	-
Interval Takeoff	48	59	58	67	TACS/JFIRE Procedures	-	-	-	-
Instrument Trail	64	72	72	80	Air to Gnd 2-Ship Mutual Supt	-	-	-	-
G-Warmup / Awareness	14	59	58	77	Mission Management	71	85	84	97
Lead Platform	33	48	46	58	VFR Arrival	68	86	80	86
Pitchout (Both)	14	48	42	57	Tanker Procedures	-	-	-	-
Fingertip (Wing)	14	54	50	68	Reciever Procedures	-	-	-	-
Route (Wing)	42	54	53	72	Airdrop Procedures	-	-	-	-
Fighting Wing (Wing)	52	57.5	58	72	Crew Coordination	63	85	80	87
Straight Ahead Rejoin	40	43	50	67	Single Engine Approach	-	-	-	-
Turning Rejoin	51	55	57	66	Single Engine GA/Missed Appch	-	-	-	-
Overshoot	44	49	49	54	A/R Overrun	-	-	-	-
Echelon (Wing)	50	50	50	50	A/R Breakaway	-	-	-	-
Breakout (Wing)	33	53	53	64	FD/AP Operations	-	-	-	-
Lost Wingman (Both)	43	62.5	59	77	FMS Operations	72	72	78	90

Table 6. Average Instructor Pilot Flight Plan Development Trends by Training Exercise

Training Exercise	# Total Events Used	# Current Events Used	# New Events Introduced	# Used Events Performed
1	19.7	19.7	19.7	0
2	26	21.9	6.3	15.6
3	29.2	21.7	3.2	18.5
4	31.1	22.3	1.9	20.4
5	33.6	23.1	2.6	20.5
6	38	23.5	4.4	19.1
7	40.6	23.1	2.6	20.5
8	44.1	24.9	3.6	21.4
9	47.8	24.9	3.7	21.2
10	51.3	24.8	3.4	21.4
11	52.9	26.6	1.7	24.9
12	54.6	25.9	1.6	24.3
13	56.8	25.5	2.2	23.3
14	57.6	23.9	0.8	23.1
15	58.3	24	0.7	23.3
16	58.4	22.6	0.1	22.6
17	58.4	21.4	0.1	21.4
18	58.7	21.4	0.2	21.2
19	59.1	20.8	0.4	20.4
20	59.3	21.8	0.2	21.6
21	59.3	22.6	0.1	22.6
22	59.5	23.3	0.2	23.2
23	59.7	22.8	0.2	22.6
24	59.8	23.1	0.1	23
25	60.1	22.8	0.3	22.5
26	60.2	24.7	0.2	24.6
27	60.3	21.1	0.1	20.9
28	60.5	25.9	0.2	25.8
29	60.6	25.2	0.1	25.1
30	60.7	26.5	0.1	26.4
31	60.8	27.8	0.1	27.7
32	61	27.6	0.2	27.3
33	61.2	26.6	0.2	26.4
34	61.2	26	0	26
35	61.4	27.8	0.3	27.6
36	61.9	25.9	0.4	25.5
37	62.3	26.6	0.4	26.2
38	62.9	24.2	0.7	23.5
39	63.6	25.2	0.6	24.6
40	64.1	25.6	0.5	25.1
41	64.9	27.6	0.9	26.7
42	65.3	23.8	0.4	23.4
43	65.5	25.3	0.2	25.1
44	65.8	26.6	0.3	26.3
45	66.8	26	1.1	24.9
46	67.8	25	1	24
47	69.2	25.2	1.3	23.9
48	70.7	27.1	1.6	25.5
49	71.9	25.8	1.2	24.7
50	72.8	27.1	0.9	26.2
51	74.1	27.6	1.3	26.3
52	74.7	29.5	0.6	28.9
53	75.4	31.2	0.8	30.4
54	77	28.4	1.6	26.9
55	77.4	29.8	0.4	29.4
56	77.9	30.2	0.5	29.7
57	79.3	26.5	1.3	25.2
58	80.3	27	1	26
59	81.4	25.5	1.1	24.4
60	81.6	28.3	0.2	28.2
61	83.3	26.2	1.5	24.7
62	83.5	26.5	0.2	26.2
63	83.9	26.9	0.4	26.6
64	85.2	27.9	1.3	26.6
65	86.2	26.7	1	25.7
66	86.8	30.5	0.6	29.9
67	87.8	28.1	0.4	27.6
68	88.6	29.2	0.8	28.5
69	89.1	27.2	0.5	26.7
70	91.3	27.8	0.7	27.1
71	92.9	31.4	1.6	29.9
72	93.8	27.3	0.9	26.4
73	94.5	27.4	0.7	26.6
74	94.8	29.5	0.3	29.2
75	94.9	25.3	0.1	25.1
76	95.6	27.6	0.6	27
77	96.5	25.4	0.9	24.4
78	95.8	26.2	0.4	25.8
79	96.6	24.8	0.8	24.1
80	97.2	22.1	0.2	21.8
81	96.5	26.6	0.3	26.3
82	96.8	24.7	0.3	24.4
83	97	21.2	0.2	21
84	97.1	22.6	0.1	22.5
85	97.3	23.6	0.2	23.4
86	97.8	21.4	0.5	20.9
87	96.2	23.1	0.1	23
88	94.2	21.2	0	21.2
89	94.4	21.8	0.1	21.6
90	94.4	21.6	0	21.6
91	92.4	23	0.1	22.9
92	92.7	23.8	0	23.8
93	93.4	22.4	0	22.4
94	94.2	22.2	0	22.2
95	99	24	0	24
96	88	27	0	27
97	88	28	0	28
98	88	22	0	22
99	88	21	0	21
100	88	16	0	16

Table 7. Average Instructor Pilot Flight Plan Development Trends by Training Exercise

Training Exercise	New From Past 1	Deleted From Past 1	New From Past 2	Deleted From Past 2	New From Past 3	Deleted From Past 3
1	-	-	-	-	-	-
2	6.3	4.1	-	-	-	-
3	4.8	5.1	3.2	7.5	-	-
4	4.6	3.9	2.8	7.2	1.9	8.7
5	5.7	4.9	4	7.2	2.9	9.4
6	8.3	7.8	5.9	10.4	5.4	12.2
7	7.8	8.2	4.3	12.6	3.7	14.5
8	8.7	6.8	5.6	11.9	4.6	15.3
9	8.4	8.4	6.6	13.4	5.4	17.3
10	8.2	8.3	6	14.6	5.2	18.7
11	9.2	7.4	6	12.6	3.2	16
12	7.4	8.1	5.9	14	4.3	17.6
13	8	8.4	5.7	14.1	4.7	19.1
14	5.4	6.9	4.2	14.2	3.5	19.2
15	7.3	7.2	5.6	12.5	3.6	17.7
16	6.2	7.6	4.3	12.9	2.9	16.8
17	5.4	6.6	3.1	11.8	2.2	16.2
18	5.5	5.6	4.1	10.7	2.9	14.8
19	5.8	6.4	3.7	9.8	3.2	14.4
20	5.7	4.7	3.5	8.9	3	11.9
21	5.7	4.9	4.4	8.3	3.6	11.7
22	6.4	5.7	4.2	8.4	3.6	11.2
23	5.6	6.2	3.7	9.9	2.7	11.7
24	6.1	5.8	3.8	9.7	2.8	12.4
25	5.6	5.9	3.4	9.5	2.8	12.8
26	7.9	5.9	5	8.9	4.7	12.2
27	4.9	8.6	2.9	12.5	2.3	14.9
28	8.8	3.9	4.2	7.9	3.5	11.1
29	6.6	7.3	4.5	9.1	3.4	12.1
30	7.2	5.9	3.8	9.8	3.4	11.3
31	7.7	6.4	4.9	9.5	3.5	12
32	5.9	6.1	3.8	10.4	2.8	12.5
33	5.8	6.8	3.4	10.6	2.7	14.1
34	7.2	7.8	4.2	11.6	2.4	13.6
35	8.3	6.4	5.2	11.1	3.9	13.7
36	5.3	7.2	3.4	11.7	2.5	15.5
37	8.1	7.4	4.4	11	3.6	14.6
38	5.6	8	3	12.8	2.3	15.8
39	8.2	7.1	5	11.9	3.7	15.5
40	7.2	6.8	4.1	10.8	3.4	14.9
41	8.2	6.2	6.3	11.1	4.7	13.5
42	6.7	10.4	4.3	14.3	3.7	18.5
43	7.3	5.8	3.8	12.8	3.1	15.9
44	7.2	5.8	4.4	8.9	2.4	13.9
45	6.8	7.4	5.4	11.8	4.7	14.2
46	8.2	9.2	4.6	13	3.8	16.6
47	7.1	6.8	4.6	13.6	3.7	16.4
48	10.4	8.6	7.9	12.9	5.5	17.3
49	7	8.2	3.3	13.2	2.9	17
50	8.1	6.8	6	12.9	3.6	15.4
51	9.4	9	7.7	14.1	5.5	18
52	8.2	6.3	4.6	11.6	2.6	14.8
53	9.3	7.7	6.3	10.9	4.8	14.8
54	7.1	9.8	4.5	14.9	3.8	17.4
55	11.1	9.7	5.2	13.6	2.6	16.1
56	9.3	8.9	5.4	14.7	4.3	17.5
57	7.4	11.2	4.4	17.1	3.5	21.9
58	7.3	6.8	3.6	14.3	3.4	20.1
59	8.3	9.8	5.9	14.2	4.7	20.5
60	8.4	5.6	5.1	12.1	4.4	15.8
61	7.2	9.5	6	14.2	4.8	19.1
62	7.8	7.5	5.4	14.6	3.8	17.7
63	8	7.5	6.2	13.3	4.5	18.7
64	8.4	7.4	5.6	12.1	4.4	16.7
65	8.9	10.2	7	15.6	5.1	18.5
66	10.5	6.7	5.9	12.3	4.4	16.1
67	9.7	12	5.9	15	3.4	18.2
68	8.8	7.6	3.8	14.6	3.3	17.1
69	7.3	9.4	4.6	14.3	2.8	19.4
70	10	9.3	6	14.5	4.6	18
71	10.9	7.3	6.5	12.1	4.5	15.4
72	6.1	10.3	4.4	15.9	3.1	19.4
73	8.6	8.5	4.9	15.1	3.4	19.1
74	9.2	7.1	5.1	11.5	2.9	15.8
75	5.8	10	2.3	13.6	1.9	17.6
76	8.2	5.9	3.1	10.7	3	14.2
77	6.8	9.1	5.1	13.3	3.9	16.9
78	6.9	6.7	2.8	12.2	2.1	15.8
79	4.7	6.1	3.7	11.8	2.8	16.4
80	3.7	6.8	2.7	12	2.3	16.5
81	8.4	3.7	6.3	6.8	4.6	11
82	4	5.9	3	8.6	3	11.7
83	2.3	5.8	2.3	11.7	1.8	13.9
84	3.4	2	2.2	6.6	2	12.3
85	4.9	3.9	4	5	3.8	9.4
86	3.8	6	2.4	8.5	1.9	9.1
87	5.6	3.7	3.7	7.9	3	9.9
88	3.1	5.4	1.8	7.9	1.4	10.2
89	3.1	2.6	2.8	7.6	2.1	9.5
90	3.5	3.6	3	5.8	2	9.8
91	3.4	1.7	2.7	4	1.9	5.6
92	3.2	2.5	2.3	3.5	1.8	5.3
93	1.6	3	1.2	5	0.8	5.8
94	1.8	1.5	0.8	3.8	0.5	5.8
95	3	0.5	2	2	1.5	3.5
96	2	0	2	0	2	0
97	1	0	1	0	1	0
98	2	8	2	8	2	8
99	1	2	0	9	0	9
100	0	5	0	7	0	14

V. Modeling Approach

5.1 Overview

Data trends combined with methods from related research are applied to design an algorithm and model that generates event sets for the next flight in a student's training campaign. This section discusses the methodology and underlying algorithms used to make these event recommendations.

5.2 Filtering Models

The flight planning process contains many components that must be accounted for to properly progress a pilot trainee through training. Volatile factors include event familiarity, individual event experience, progress through the training program, and progress in comparison to the rest of the class; all factors are influential in making flight planning recommendations.

Predicting accurate, personalized recommendations for events that fit a pilot training student's true progression and challenge the student's abilities requires a complex solution. Fitting a hybrid model composed of both content-based and collaborative filtering is proposed as the best approach for such a complex problem.

At each training exercise, a recommendation is made for the next set of flight events in a continuous sequence of event sets that make up a student's training history. Event set recommendation is viewed as an advanced sequence prediction task called a sequential set-to-set task. The fundamental sequence prediction task aims to predict the next value in a sequence based on the existing values in a sequence. Sequence prediction has been used on a variety of tasks such as predicting price value changes based on temporal price trends or predicting the next alphabet character in a computer generated sentence based on all of the characters that preceded it.

Flight recommendation is performed similarly. However, at each interval, the model produces a set of values rather than a single value. Each value in the set represents an individual event recommended for the next flight plan.

5.2.1 Motivation for Content-Based Model

Sequence prediction is a task that involves using temporal sequence data to predict the next value or values in the sequence. For PTN, sequence prediction uses the previous flight data of an individual student, such as the progression of accomplished events and corresponding grades, to recommend a vector of events appropriate for the next flight plan. Industry-leading models for dealing with this type of sequential prediction are long short-term memory recurrent neural networks (LSTM RNNs) and temporal convolutional networks (TCNs).

Deep learning models learn from the data provided, so limited data records may result in limited output options for more well established models such as LSTM RNNs or TCNs. PTN has very limited records of student data due to it's relatively short history. Limited training data may lead to a limited variety of output recommendation options from LSTM and TCN models, so the outputs are subject to standardization in length and lack personalization in flight event combinations.

The proposed solution to combat the limitations of output vectors resulting from standard sequence prediction models is sequential set generation. Each unique flight plan in undergraduate pilot training consists of a set of individual flight events, not necessarily a specific vector of events that must be paired together every time. Sequential set generation iteratively outputs individual elements of a set of events rather than a vector with a fixed number of flight events. Iteratively generating each individual component of each set of flight plan events allows for new, more personalized recommendations for the individual training exercise that avoids the limitations of standardized vector outputs and more accurately represents how IPs build flight plans.

5.2.2 Motivation for Collaborative Models

Collaborative filtering is used to incorporate a competitive aspect between a pilot training student's progress and other students' progress throughout training. A student's performance at each training level is assessed and compared to that of their peers to encourage continuous performance advancement for individuals and succeeding pilot training classes. Incorporating collaborative progress into recommendations is a preventative measure to avoid individual students from falling behind in training due to a slower natural rate of advancement compared to pilot training peers.

Setting a golden standard for performance provides a baseline for making recommendation adjustments that converge towards a global performance expectation. Establishing a global baseline assures that all students are progressing in a similar manner as the top performers, even if at a slower pace.

Combining collaborative filtering models with content-based filtering models using deep learning architectures introduces the ability to make more personalized recommendations for the user. Two different content-based sequential structures for content-based recommendation using deep learning provide adequate methods of approaching the AutoGradebook event recommendation process: sequence prediction and sequential set generation.

5.2.3 Motivation for Hybrid Model

The model uses an ensemble of content-based and collaborative filtering methods. A content-based model provides more personalized recommendations for each student derived from personal training and performance. The PTN program provides for more personalized training while also graduating pilot training students in a fraction of the time that the current UPT program takes. Personalized training may promote effective training progress, but providing recommendations solely on a particular student's performance history may result in periods of stagnated growth and ultimately longer training duration in the program. Collaborative filtering helps limit stagnation by establishing training expectations from the global population. As a stochastic variable, student improvement cannot always be predicted accurately. Continuously providing recommendations that guide individual student performances towards a global performance expectation is one method of combating such stochasticity.

5.3 Proposed Content-Based Model

Literature has shown that multiple deep learning model structures can be used to adequately make predictions in sequence to sequence tasks. The current most commonly applied deep learning model for sequence prediction seems to be the LSTM RNN. An LSTM RNN was applied in this research to make initial personalized event set recommendations given temporal training data. Model construction and fitting were conducted via the TensorFlow software library with the Keras wrapper within the Python modeling environment. Exploration of alternative model structures and applications is proposed for future research.

5.3.1 Model Architecture

The content-based model architecture is composed of two stacked LSTM RNN layers. Each layer uses an activation function to take inputs and produce outputs that will be used as inputs for the next layer in the network. A nonlinear activation function is used given the complexity of the problem and data patterns. Each LSTM layer uses a rectified linear unit (ReLU) activation function. The ReLU activation allows network parameters to converge toward optimum quickly with the use of backpropogation. ReLU activation functions also help avoid vanishing gradients, a problem that often leads to ineffective parameter updates during the backward propagation training process and a less effective model. The output layer uses a sigmoid activation function to convert all predicted values to a value between 0 and 1. Flight events with predicted values greater than or equal to 0.5 are included in the content-based recommendation. The model consists of 184,928 trainable parameters. Model structure is represented in Figure 8.

5.3.2 Data Preparation

The model inputs consist of student scores recorded as percentages of the individual event MIF required for proficiency. Inputting scores as a percentage of proficiency normalizes the values, allowing for more efficient parameter training. Sequential data must be structured properly to input it into an LSTM RNN model. Model input sequence lengths of 10, 25, and 50 were explored for model performance purposes. All input sequences must be the same size when training a model. Given the temporal nature of evaluation history, it may not be possible to create an input sample with a certain number of sequential data records. For example, it is not possible to make an input sequence of length 50 if a student has not conducted at least 50 training exercises. To counter this issue, all missing sets were forward padded with sets of 0 values representing training exercises with no flight event evaluations. Forward padding the inputs enables model training for a variety of evaluation history quantities while maintaining consistent input size. Only the most recent training exercise evaluations, with a total number of evaluations equal to the sequence length, were included in the input if the input correlated to a training exercise recommendation beyond the designated sequence lengths.

Target variables sets were also restructured to create an appropriate training and

```
Model: "Content-Based LSTM RNN Model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
LSTM_Layer_1 (LSTM)          (None, 50, 100)           91600
_________________________________________________________________
LSTM_Layer_2 (LSTM)          (None, 100)               80400
_________________________________________________________________
Output_Layer (Dense)         (None, 128)               12928
=================================================================
Total params: 184,928
Trainable params: 184,928
Non-trainable params: 0
_________________________________________________________________
```

Figure 8. Long Short Term Memory Recurrent Neural Network Model Architecture

testing environment for the model. The training targets, referring to the single set of flight events that the model compares results with during training to make parameter adjustments, were gathered from the succeeding training event of the input sequence. The training targets were reformatted from student grades to binary representations of event occurrence.

5.3.3 Model Fitting

A supervised learning approach was applied to train the model using classification accuracy and binary cross-entropy loss metrics. Accuracy measures how similar the models recommendations are to the true IP recommendations in the data. This metric is calculated as the percent of events in the content-based recommendation that were correctly listed as occurring or not occurring when compared to the true IP recommendations. The equation to measure accuracy between the real IP recommendations and model generated recommendations is shown in Equation 1, where b is a binary variable representing whether or not the target and predicted variables are equal and e represents the specific graded event. b_e equals 1 if the model's predicted value rounded to the nearest whole number is equal to the true IP recommendation for each flight event, e, and equals 0 otherwise.

$$Accuracy = -\frac{1}{128}\sum_{e=1}^{128} b_e \tag{1}$$

Whether or not the model outputs the correct suggestion for all 128 graded events is incorporated into this metric because an IP must make a binary decision whether to include each graded event into the flight plan for each new training event. Binary cross entropy measures the confidence of the recommendations using Equation 2, where t represents the target value chosen by a real IP, p represents the models predicted value, and e represents the specific graded event.

$$BCE(t_e, p_e) = -\frac{1}{128}\sum_{e=1}^{128} t_e * log(p_e) + 1 - t_e * log(1 - p_e) \tag{2}$$

The sigmoid activation function in the output layer uses a value between 0 and 1 to classify events as occurring or not, respectively. If the model outputs the value of 0.5 for an individual event, that means the model's confidence is evenly split between the two classification. Values closer to 1 or 0 represent more confident model predictions. Lower binary cross-entropy represents that the sigmoid function is generating values closer to the real binary recommendation values of 0 or 1. Model accuracy did not make drastic increases with the addition of more training epochs, so training completion was determined by digression of the loss metric.

High personalization of individual student training campaigns means that each students records are independent of the others. Students 1-4, which represent a variety of training durations, were set aside for a model testing set. The remaining 14 students composed a training set used to perform leave-one-out cross validation (LOOCV) to find a best-fit model. The grade sheet data were restructured as three dimensional arrays with a designated sequence length number of training exercise evaluations that acted as input for the model. A vector of binary values, 1 designating that an event was evaluated and 0 designating otherwise, acted as the target variable associated with each input. Parameter optimization was performed using three different sequence lengths to check for significance in the amount of prior evaluation history needed to effectively provide a recommendation.

Model parameter optimization was performed using 13 of the training examples and then performance was validated on the remaining training examples. The last training record from each student was not used in model training due to the lack of an available target set. A single training epoch refers to the complete presentation of 13 training examples to the model. The model was arbitrarily set to train for 100 epochs, meaning the data may continuously be fed into the model up to 100 times to make parameter adjustments. Applying an early stopping technique prevented the model from continuing to train if the loss metric did not improve after 15 epochs. Binary cross-entropy improvement was defined as a decrease of less than 0.001. This process was repeated 14 separate times to create 14 different models with corresponding accuracy and loss metrics. The model with the lowest validation loss measurement was chosen as the final model. The LOOCV validation technique was chosen for this task to maximize the number of training examples used for parameter optimization given data limitations from only 18 total students. Using more data in training helps fit a model that can generalize predictions to the larger PTN student population rather than overfitting a model to any specific training path. Figure 9 and Figure 10 show the change in model accuracy and loss binary cross entropy at each training epoch, respectively. Validation accuracy does not improve much throughout training. With only 13 students, limited data may be preventing further accuracy improvements. Validation loss improved drastically during the first few training epochs, but leveled out around epoch 10. The model continued training until a loss value improvement of 0.001 did not occur for 15 consecutive training epochs. The model parameters at the epoch with the best loss value were recovered and used as the best fit content-based model.

Table 8 shows the results of the best fit models trained using each sequence length. Small variation in result values was observed from the models created in the LOOCV process for each respective sequence length. Small variances are likely due the sparse training data available. Further parameter tuning was discontinued due to lack of data and results variety.

Table 8. Leave-One-Out Cross Validation Model Results

Sequence Length	Training Loss	Validation Loss	Training Accuracy	Validation Accuracy
50	0.1759	0.1771	0.9242	0.9241
25	0.1780	0.1682	0.9241	0.9247
10	0.1886	0.1713	0.9238	0.9266

5.4 Proposed Collaborative Model

5.4.1 Establishing Performance Standards

The golden standard is defined using the total event exposure and average recorded grades of the top ten percent of student records in the PTN database for the given training exercise number. On the given training exercise, the average MIF ratio score of the top ten percent of global performers is calculated for each possible event. The set of average MIF ratio values makes up the golden standard for performance at the given training exercise. This set of values provides a reference for expected

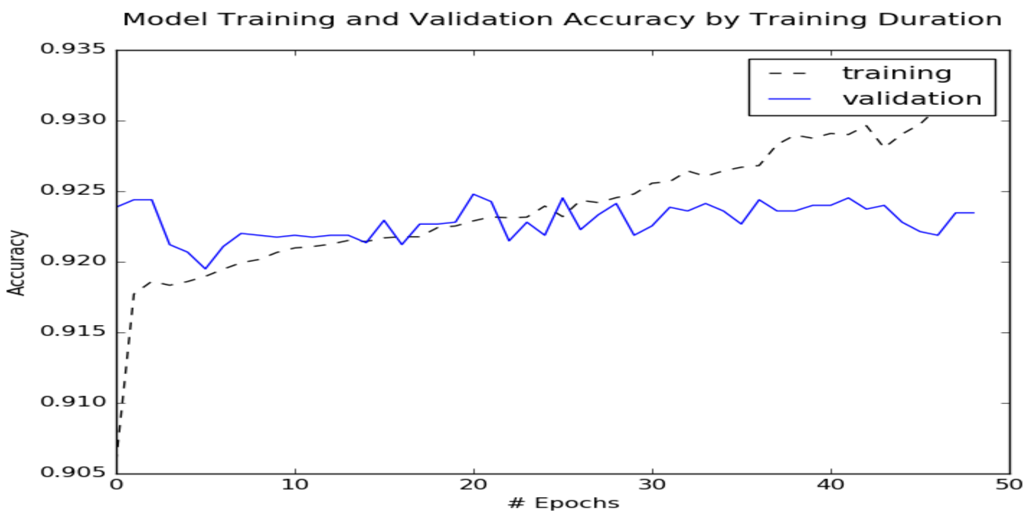


Figure 9. Training and Validation Accuracy per Epoch

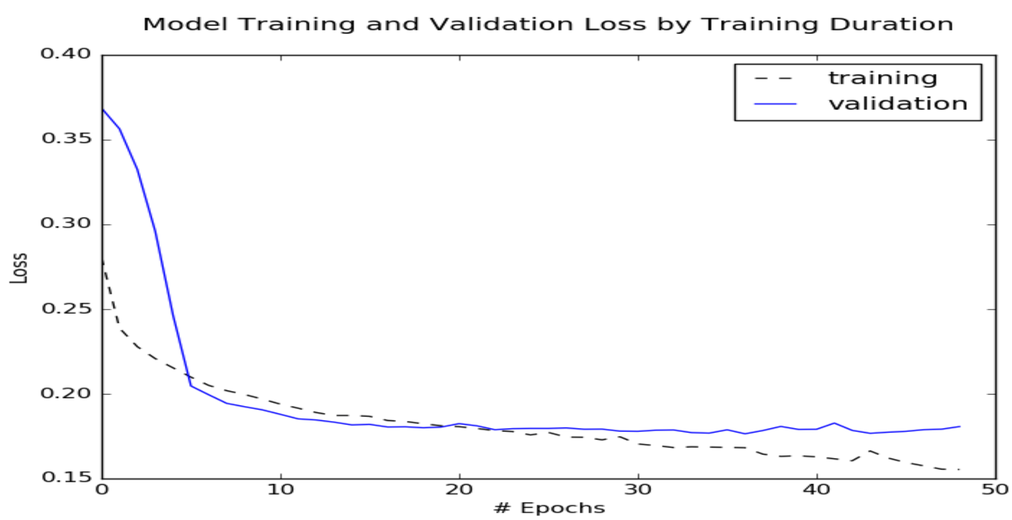


Figure 10. Training and Validation Loss per Epoch

progress towards proficiency in each graded event upon reaching the given training exercise. The FPS value for the golden standard is calculated as a reference for student orientation around the established progress expectation.

An individual event progress (IEP) score is calculated for each event by subtracting a student's individual MIF ratio value from the golden standard MIF ratio values. The resulting IEP value represents how far a student is from the expected progress standard. A positive value means the student's progress for a specific event being below the golden standard for the event at the given training exercise. A negative value means the student's progress for a specific event is above the golden standard at the time. A student is deemed above the golden standard if their set of IEP scores does not contain any positive numbers, indicating they are above the standard for every event. Otherwise, a student is deemed below the golden standard.

Collaborative filtering for recommendations is tailored based on each student's orientation to the golden standard at the given training exercise. Recommendations given to any student resting below the establish golden standard are tailored to progress them towards the standard. Recommendations for students resting above the golden standard are tailored strictly to improve student FPS.

5.4.2 Below the Standard

Inspired by the PSO algorithm, the collaborative filtering model compares the current student progress scores of each possible event with the scores in the golden standard. Collaborative based recommendations for students below the golden standard are tailored to advance students towards the golden standard using event specific adjustments similarly to how the PSO algorithm adjusts point conditions toward a population optima.

The set of IEP scores is implemented along with the likelihood of the specific event occurring at the given point in training to calculate the utility of recommending a specific event on the given training exercise. The event with the most positive utility provides the most benefit to student progress and has the most incentive for recommendation. The event with the most negative utility provides the most benefit to student progress and has the least incentive for recommendation. All event utilities are calculated and stored in descending ordered.

The model then separately scans all events recommended from the content-based model and the list of all events excluded from the content-based recommendation. The event with the lowest utility from the initial recommendation is swapped out with the event with the highest utility from the list of initially excluded events. No event swapping will take place if all events from the content-based recommendation yield higher utility than any of the excluded events.

A student below the standard in many graded events may require more tailored recommendations to push them towards the expected skill level for a given training exercise. Therefore, the algorithm makes event swaps accordingly using the values found in Table 9. The ranges for number of events in Table 9 were set arbitrarily and only allow fewer swaps to be performed than the number of events below the golden standard. No more than 6 event swaps from the content-based recommendation may be performed. Thus, the influence of the collaborative model on final recommendations is limited to maintain the personalization of training campaigns gained from the content-based model.

Incorporating additional performance standards, such as a local standard, for more tailored recommendations has been explored for the given problem, as well. This application proved unnecessary current task, but may be useful in alternative applications of the AutoGradebook. Further discussion on the incorporation of additional performance standards can be found in the Future Work section.

Table 9. Number of event swaps according to the number of graded events below the golden standard (GS)

Number of events below the GS	Event swaps preformed
1-5	1
6-10	2
11-20	3
21-30	4
31-40	5
41+	6

5.4.3 Above the Standard

Students above the golden standard are still expected to improve throughout training, so the flight recommendation system must continue to suggest events that will improve student progress. However, working toward the golden standard no longer benefits student progress in this case. Unlike the PSO algorithm, this model does not attempt to redirect students back toward the golden standard once they have passed it. Instead, a separate greedy heuristic search algorithm is applied to enhance student learning and performance beyond the golden standard. Using the FPS metric provides an alternative approach to make recommendations that are predicted to maintain student training progression.

The FPS is calculated using a hierarchical point system, incorporating three components of student progression to more precisely express student advancement through training. The three components of student progression determining the FPS include individual event scores, introduction to new event categories, and proven proficiency in event categories. Depending on student progress, it is an IP's subjective opinion whether to choose to introduce new events rather than emphasize on events where students have already received graded evaluations and vice versa. Point allocations for each component of the FPS are derived according to the preference of each component of progress. The preferences of the student progress components are expressed in descending order in Table 10.

Table 10. Ranking of Possible Progressive Events Occurring From Recommendation

Ranking	Event
1	Achieving proficiency on all events in a category
2	Achieving proficiency in a single event
3	Introduction of all events in a category
4	Introduction of a single event
5	Individual event performance increase from U to F
6	Individual event performance increase from F to G

Events in descending order of value

Complete proficiency in all events is the overall goal of each student training campaign. Therefore, achieving proficiency across multiple events is more valuable than achieving proficiency in one event and weighted as the most valued event that can occur. An event must be introduced for student progress in that event to improve. Achieving proficiency in a single event is the next most favorable outcome of a training event.

Achieving near-proficiency across multiple events is more beneficial to the overall training campaign than achieving near proficiency in a single event. It is not possible to achieve near proficiency in multiple events if only a single event has been introduced. Therefore, points allocated for being introduced to a set of events are greater than points allocated for an event evaluation that is only one letter grade away from proficiency. Naturally, it is more valuable for multiple events to be introduced than for a single event to be introduced.

Introducing a new event shows progress in both skill depth and breadth, so it is more valuable for a student to be introduced to a new event than for the student to progress without reaching proficiency. In addition, the data shows that a good score is the most frequent evaluation score for students. This indicates that progress tends to stagnate just before achieving proficiency and students are expected to maintain a good evaluation score over multiple training exercises rather than immediate progression to proficiency. To account for this stagnation, progressing from a fair grade to a good grade is the least valuable progressive event outcome from a training exercise.

The first component of the FPS calculation applies the individual event scores to account for training depth. Each event has its own individually designated MIF score to represent proficiency in that event. Given that each event may have a different individual MIF score, taking the cumulative MIF score introduces more ambiguity regarding how close to proficiency a student may be. Calculating the fraction of a student's current maximum score over the designated individual MIF score for each individual event provides a clearer metric of a student's progress toward proficiency in an individual event. The individual event fraction (IEF) is calculated using Equation 3.

$$IEF_{man} = \frac{MaxScore_{man}}{MaxMIF_{man}} \tag{3}$$

In Equation 3, man represents an individual flight event index, $MaxScore_{man}$ represents the student's highest earned score during any graded evaluation of event man, and $MaxMIF_{man}$ represents the minimum score a student must receive on event man in order to be deemed proficient in that event. IEF_{man} represents the student's calculated IEF score for the specified event, man.

Exploratory data analysis concluded that recorded evaluation data seemed to contain IP bias that influenced how quickly a student advanced their individual event performance. The IEF directly incorporates the biased data. A model using biased data must account for that bias in order to generate a more appropriate prediction. Therefore, a hierarchical point system was developed in accordance with trends from provided data to allot points for student performance on each event based on IEF values. Maximum individual MIF scores for all events are either 3 or 4, so divisional boundaries within the hierarchy were established with consideration that not all maximum MIF scores are equal. Receiving a fair score in any event with a maximum MIF score of 3 receives equivalent progress points to receiving a good score in any event with a maximum MIF score of 4 to simplify model calculations.

Incorporating IEF into the FPS accounts for the depth component of student progress. For each event, the Cumulative Event Points (CEP) scores are allotted based on IEF and inserted into the FPS equation. The value hierarchy of allotted points is expressed in Table 11.

Table 11. Point Allocation Per Event Graded Evaluation

IEF value	Additional Points Allotted	Cumulative Points Allotted
$IEF \geq 1$	2.25	5
$1 > IEF \geq \frac{2}{3}$	0.50	2.75
$\frac{2}{3} > IEF > \frac{1}{3}$	0.75	2.25
$\frac{1}{3} \geq IEF > 0$	1.5	1.5
$IEF = 0$	0	0

The hierarchical point system is not collective. This means that for each event a student can only receive one of the values from the Cumulative Points Allotted column of Table 11. The possible CEP values for each event are expressed in the Cumulative Points Allotted column of Table 11. A student will never have a CEP score greater than 5 for an individual event throughout a training campaign. Applying the points system defined in Table 11, the equation to calculate the first component of the FPS score is expressed as:

$$Component1 = \sum_{man \in manPossible} CEP_{man} \tag{4}$$

where *man* represents an individual flight event and *manPossible* represents the set of all possible flight events for evaluation.

The second and third components of the FPS calculation incorporate the ten event categories to account for training breadth. The second component applies an additional point bonus to the FPS if all events within an event category have been introduced at any point up to the current training exercise. The equation to calculate the second component of the FPS score is expressed as:

$$Component2 = \sum_{cat=1}^{10} SWeight_{cat} * S_{cat} \quad \forall \quad SWeight \in SeenWeights \tag{5}$$

where *cat* represents an individual event category index, *SWeight* represents the defined weighted value of being introduced to all events within a category, *cat*, by the given training exercise, and S_{cat} is a binary variable representing whether all events within a category, *cat*, have been introduced to the student by the given training exercise. *SWeights* represents the set of individual weighted values corresponding to the introduction of all events for each event category. A value of 1 for the variable S_{cat} means that all events within a category, *cat*, have been introduced to the student, 0 otherwise.

The third component applies an additional point bonus to the FPS if student performance in all events within an event category reach proficiency any point up to the current training exercise. The equation to calculate the third component of the

FPS score is expressed as:

$$Component3 = \sum_{cat=1}^{10} MWeight_{cat} * M_{cat} \quad \forall \quad MWeight \in MIFWeights \tag{6}$$

where *cat* represents an individual event category index, *MWeight* represents the defined weighted value of reaching proficiency in all events within a category, *cat*, by the given training exercise, and M_{cat} is a binary variable representing whether all events within a category, *cat*, have reached proficiency to the student by the given training exercise. *MIFWeights* represents the set of individual weighted values corresponding to reaching proficiency in each event category. A value of 1 for the variable M_{cat} means that the student has reached proficiency in all events within a category, *cat*. A value of 0 for the variable M_{cat} means that the student has not reached proficiency in all events within a category, *cat*.

The final equation for the calculation of the FPS for a single student is:

$$FPS_{stud,TE} = Component1 + Component2 + Component3 \tag{7}$$

where *stud* represents a specific student and *TE* represents the training exercise having the FPS.

A set of advancing event combinations is produced by implementing swaps between the content-based recommendation and the list of excluded events. Advancing combinations are created by making a one-for-one swap between an event in the content-based recommendation and an excluded event yielding higher event utility. Only one event swap is implemented because the student is already performing above the standard. There is no guaranteed way to ensure continued progress, but a single swap designed to raise the FPS adds an element of predicted advancement in event depth or breadth that is most beneficial to a student's progress state at the time.

Corresponding FPS values are calculated for each advancing event combination and the initial content-based recommendation. A greedy heuristic algorithm is applied to identify the set of events yielding the highest predicted FPS for the immediate future. The set of events yielding the highest predicted FPS constitutes the final recommendation for the user. A single random selection from the premier event sets is chosen as the final recommendation when various sets of events yield equally dominating FPS scores.

The Recommender System Algorithm shows the complete process from user inputs to final recommendations.

Recommender System Algorithm

```
User initializes userID and TE
Initialize userData as all data from userID
Initialize globalData as all data excluding userID
for each flight event do
    Calculate IEF_userID
end for
for each event category do
    Check if all events have been introduced
    Check if all events have reached proficiency
end for
Calculate current FPS score for userID
Restructure userData for LSTM inputs
Categorize event recommendations using LSTM model for initial recommendation
Calculate userID FPS for TE+1 according to initial recommendations
for each student in globalData do
    for each flight event do
        Calculate IEF
    end for
    for each event category do
        Check if all events have been introduced
        Check if all events have reached proficiency
    end for
    Calculate student FPS at TE
end for
Identify top 10% performers at TE based on maximum FPS
for each flight event do
    IEF_GS = Avg(IEFs in top 10%)
    Calculate P(event occurring on TE
    studentStatus = (IEF_GS) − (IEF_userID)
    Utility(flight event) = P(event occurring on TE)* studentStatus
end for
if any(studentStatus)≤ 0 then
    Perform all possible utility advancing event swaps to create various alternative recommendations from the initial recommen-
    dation
    Calculate projected FPS for all alternative recommendations
    Final Recommendation = Initial Alternative Recommendation with Max Projected FPS
else
    Make 1-for-1 event swapping adjustments from initial recommendation
    Final Recommendation = adjusted initial recommendation
end if
```

The Recommender System Algorithm begins by initiating the user ID and next training exercise in training. Next, grade sheet data is split between into two data sets:

the user's evaluation history and a global data set excluding the user's history. The user's IEF values are then calculated for each flight event and each category is checked for introduction and full proficiency of all events. FPS scores are calculated for each student at the initialized training exercise number. Next, the user's evaluation data is restructured into a three dimensional array and input into the content-based LSTM RNN model to generate an initial recommendation. Lines 15-24 in the Recommender System Algorithm calculate the current FPS for all students in the Global data. Next, the top 10% of performers in the global data are identified and the average IEF values between them define a golden standard for training performance. Lines 26-31 define how far away the user is from reaching proficiency in each maneuver and calculate an expected utility for recommending each event. The user is then labelled as above or below the golden performance standard. Lines 33-35 show that if the user is above the golden standard, then utility-advancing alternatives to the initial recommendation are generated and the option with the highest projected FPS score is used as the final recommendation for the user. Lines 36-38 show that if the user is below the golden standard, then event adjustments will be made to the initial recommendation in an attempt to progress the user towards the golden standard. The adjusted recommendation is then used as the final recommendation.

VI. Results and Analysis

6.1 Overview

This section presents final testing results for the content-based model. Also, final results produced by the full hybrid recommender system are explained and analyzed.

6.2 Model Results

6.2.1 Examples of Recommended Flight Event Sets

Figures 11 and 12 are examples of the evaluation data used to plan a student's next training exercise, the flight plan the IP put together for the student's next training exercise, and the flight plan that the model recommended for the student's next training exercise. Figure 11 shows the model's recommendation compared to an IP's actual flight plan for a student's third training exercise. As shown, only data from the first two training events is available to help make a decision for the next flight's graded events. Figure 12 shows the model's recommendation compared to an IP's actual flight plan for the same student's twenty-first training exercise. In Figure 12, there are 20 previously evaluated training exercises available to help make a decision for the next flight's graded events. While IPs use all of a student's evaluation data, the model only takes up to 50 of a student's most recent training exercise evaluations into consideration when making recommendations.

6.2.2 Content-Based Recommendation Testing

The content-based model creates initial recommendations based on IP generated data. The incorporation of the collaborative model is designed to make improvements to content-based recommendations using global trends in the data. Therefore, it is

Training Exercise	1	Score	2	Score	3	3 (IP-Chosen)	3 (Model-Generated)
Graded Events	Mission Analysis/Products	F	Mission Analysis/Products	G	?	Mission Analysis/Products	Mission Analysis/Products
	Ground Ops	F	Ground Ops	F		Ground Ops	Ground Ops
	Takeoff	G	Takeoff	F		Takeoff	Takeoff
	Departure	U	Departure	U		Departure	Departure
	Basic Aircraft Control	F	Basic Aircraft Control	F		Basic Aircraft Control	Basic Aircraft Control
	Cross-Check	U	Cross-Check	F		Cross-Check	Cross-Check
	Enroute Descent / Recovery	F	Enroute Descent / Recovery	G		Enroute Descent / Recovery	Enroute Descent / Recovery
	Inflight Checks	U	Inflight Checks	U		Inflight Checks	Inflight Checks
	Inflight Planning	U	Inflight Planning	F		Inflight Planning	Inflight Planning
	Communication	U	Clearing / Visual Lookout	U		Clearing / Visual Lookout	Clearing / Visual Lookout
	Risk Mgmt / Decision Making	U	Communication	F		Communication	Communication
	Situational Awareness	F	Risk Mgmt / Decision Making	U		Risk Mgmt / Decision Making	Risk Mgmt / Decision Making
	Task Management	U	Situational Awareness	U		Situational Awareness	
	Emergency Procedures	F	Task Management	F		Task Management	Task Management
	General Knowledge	F	Emergency Procedures	F		Emergency Procedures	Emergency Procedures
	Vertical S	F	General Knowledge	F		General Knowledge	General Knowledge
	Steep turns	G	Overhed/Closed Pattern	F		Overhed/Closed Pattern	Overhead/Closed Pattern
	Precision Final	F	Landing	U		Landing	Landing
			Go-Around	F		Go-Around	
			G-Awareness	F		Emergency Landing Pattern	
			TP Stalls	F			Slow Flight
			Slow Flight	F			
			Power on Stalls	F			
			Aileron Roll	F			
			Loop	F			
			Split S	F			
			Full Procedure Approach	U			
			Precision Final	F			

Figure 11. IP-chosen event recommendations vs. model-generated event recommendations on training exercise 3

Training Exercise	1	Score	2	Score	•••	19	Score	20	Score	21	21(IP-Chosen)	21 (Model-Generated)
Graded Events	Mission Analysis/Products	F	Mission Analysis/Products	G		Mission Analysis/Products	G	Mission Analysis/Products	G	?	Mission Analysis/Products	Mission Analysis/Products
	Ground Ops	F	Ground Ops	F		Takeoff	G	Ground Ops	G			Ground Ops
	Takeoff	G	Takeoff	F		Departure	F	Takeoff	G		Takeoff	Takeoff
	Departure	U	Departure	U		Basic Aircraft Control	G	Departure	G		Departure	Departure
	Basic Aircraft Control	F	Basic Aircraft Control	F		Cross-Check	F	Basic Aircraft Control	G		Basic Aircraft Control	Basic Aircraft Control
	Cross-Check	U	Cross-Check	F		Enroute Descent / Recovery	G	Cross-Check	G		Cross-Check	Cross-Check
	Enroute Descent / Recovery	F	Enroute Descent / Recovery	G		Inflight Checks	F	Enroute Descent / Recovery	G		Enroute Descent / Recovery	Enroute Descent / Recovery
	Inflight Checks	U	Inflight Checks	U		Inflight Planning	G	Inflight Checks	G		Inflight Checks	Inflight Checks
	Inflight Planning	U	Inflight Planning	F	•••	Clearing / Visual Lookout	G	Inflight Planning	G		Inflight Planning	Inflight Planning
	Communication	U	Clearing / Visual Lookout	U		Communication	F	Clearing / Visual Lookout	G		Clearing / Visual Lookout	Clearing / Visual Lookout
	Risk Mgmt / Decision Making	U	Communication	F		Risk Mgmt / Decision Making	G	Communication	F		Communication	
	Situational Awareness	F	Risk Mgmt / Decision Making	U		Situational Awareness	G	Risk Mgmt / Decision Making	G		Risk Mgmt / Decision Making	Risk Mgmt / Decision Making
	Task Management	U	Situational Awareness	U		Task Management	G	Situational Awareness	G		Situational Awareness	Situational Awareness
	Emergency Procedures	F	Task Management	F		General Knowledge	G	Task Management	G		Task Management	Task Management
	General Knowledge	F	Emergency Procedures	F		Intercept/Maintain Arc	G	Emergency Procedures	G		Emergency Procedures	Emergency Procedures
	Vertical S	F	General Knowledge	F		Full Procedure Aproach	F	General Knowledge	G		General Knowledge	General Knowledge
	Steep turns	G	Overhed/Closed Pattern	F		Non-Precision Final	G	Wing Takeoff	G			Landing
	Precision Final	F	Landing	U	•••	Circling Approach	F	Lead Platform	G		Vertical S	
			Go-Around	F				Fingertip(Wing)	G		Intercept/Maintain Arc	
			G-Awareness	F				Route(Wing)	G		Full Procedure Approach	
			TP Stalls	F				Straight Ahead Rejoin	G		Non-Precision Final	Non-Precision Final
			Slow Flight	F				Overshoot	G		Precision Final	Precision Final
			Power on Stalls	F				Echelon(Wing)	G		Circling Approach	
			Aileron Roll	F				Breakout(Wing)	G		Missed Approach	
			Loop	F				Lost Wingman(Both)	G			
			Split S	F				Extended Trail(Both)	G			
			Full Procedure Approach	U				Position Change	G			
			Precision Final	F	•••							

Figure 12. IP-chosen event recommendations vs. model-generated event recommendations on training exercise 21

valid to exclude utility advancing maneuver swaps when measuring model performance.

Testing was performed on the best-fit LSTM RNN model identified in the Model Approach Chapter using the evaluation history of the four students not included in the training set. Table 12 depicts the final testing results for the model generated using each input sequence length. The testing loss and accuracy were calculated using Equation 2 and Equation 1, respectively. Testing results showed only a fraction of a percent difference in performances between models with different sequence lengths. The final model selected uses a sequence length of 50, meaning that up to 50 previous training exercise evaluations are taken into consideration when generating each recommendation. The driving factor behind this decision was that trends in the data showed that sometimes students are introduced to a flight event early in their training campaign and then would not revisit that event until much later. Using a sequence length of 50 ensures that event progress is recognized by the model despite the time since a student last performed that event.

Table 12. Model Testing Results

Sequence Length	Testing Loss	Testing Accuracy
50	0.2146	0.9145
25	0.2113	0.9170
10	0.2151	0.9136

Figure 13 shows model testing accuracy with respect to training exercise. Volatility in accuracy with regard to training exercise number suggests that the model is not performing consistently throughout training campaigns. More specifically, the model performs more accurately before a student's fifth training exercise, between the twelfth and twentieth training exercise, and after exercise 75. The model performance suffers around the fiftieth training exercise. The reasons for this are unknown. However, high model performance suggests that the model makes flight plan recommendations similar to true IP recommendations.

6.2.3 Final Maneuver Set Recommendation

The resulting output of this hybrid model is the final recommendation for the user. The final recommendation is composed of a set of training events to perform and be evaluated during the succeeding training exercise of a students training campaign. The order of the flight events recommended has no significance to the user and does not imply an order of execution during the training exercise.

The primary driver for automating the flight planning process is to reduce the amount of time IPs spend on administrative tasks, allowing them to reallocate their time to working directly with the students. To test recommendation time, the full hybrid model was used to make recommendations for each training exercise of the students in the training set. The time of generation was recorded for flight plan recommendations generated for 331 individual training exercises. Time testing results show an average recommendation time of 8.14 seconds with a standard deviation of 0.10 seconds. The median generation time was 8.04 seconds. Generation times ranged from a minimum of 7.93 seconds to a maximum of 10.05 seconds.

6.3 Analysis of Results

Standard model training, validation, and testing techniques offer a baseline for recommender system performance expectations. However, recommender systems are often designed to improve user experience when there is no single correct path forward, so the best way to test this system is through real-time trial and error with IPs. Developing personalized training campaigns for every student in conjunction with the lack of a single optimal path towards fully qualified status makes generating recommendations with 100 percent accuracy extremely improbable. Perfect accuracy

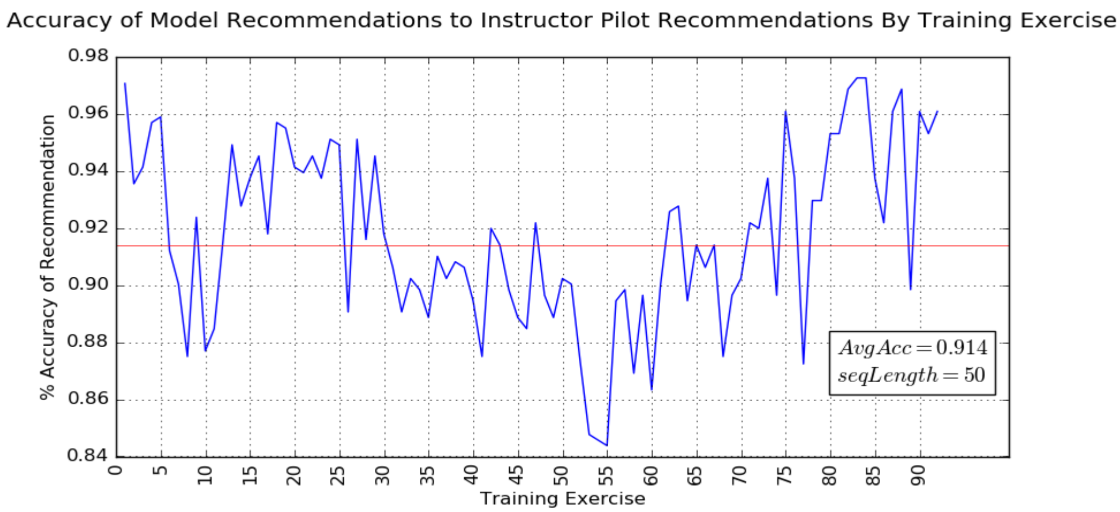


Figure 13. Model Accuracy with regard to Training Exercise

in recommendations would suggest the presence of a standardized syllabus for training progression rather than personalized recommendation tailored to individual student needs.

High model performance despite sparse data suggests that there are consistent trends for flight plan generation that are generalized and used across the students population. This also suggests that the performance dips in Figure 13 show the region of training duration where PTN IPs are incorporating the most personalization to student flight plans.

The resulting recommendation time statistics prove that automating the flight planning process has the ability to drastically decreases the amount of time an IP must allocate for flight planning from hours to only seconds on a daily basis. This tool shows potential to play a substantial role in decreasing IP time spent performing administrative responsibilities even if the IP takes a few minutes to adjust each recommendation.

VII. Conclusions

The Air Force's PTN program has had success in effectively shortening the length of undergraduate pilot training campaigns for pilot training students. Operational differences from traditional undergraduate pilot training through the emphasis of virtual reality flight simulation alongside periodic real aircraft experience allow for a more efficient pilot training process. Automating tedious tasks performed by IPs provides an opportunity to make the pilot training process even more efficient. Student evaluation data from the original pilot training class were provided from PTN to explore methods of flight plan automation and uncover program insights.

Exploratory data analysis was conducted on pilot training student evaluation data from the original PTN class to provide insights on the flight evaluation process, student progress trends, and the current data collection and storage practices of the PTN program. The data unveils nonlinear trends in skill depth advancement, allowing students to move away from lower scores quickly but stagnate just under proficiency for longer. Inconsistencies in the overall flight event and category exposure between students appears to be linked to training campaign duration, not simply student performance. Also, students appear to be graduating from the PTN program without official record of full proficiency in all graded flight events. Creating an administrative tool to aid IPs presents a low risk problem because the IPs can always make adjustments to recommendations. However, clear graduation requirements must be established and enforced to create a fully autonomous tool, such as the expectation of the AutoGradebook.

A new metric, called the Forward Progress Score (FPS), was developed to better track student progress throughout undergraduate pilot training. Unlike the method currently employed, FPS incorporates both breadth and depth of student skill advancement. The FPS uses progress tracking metrics such as proficiency of individual graded events, quantity of introduced events and event categories, and proficiency of entire event categories to capture the multidimensional aspect of student advancement through a training campaign.

A hybrid filtering approach using both content-based and collaborative models was applied to generate a set appropriate flight events for evaluation in a students next training exercise. A long short-term memory recurrent neural network was trained and tested on real IP recommendations from the data to produce an initial student-specific recommendation. Testing solidifies the model's ability to produce flight event recommendations averaging 92 percent similarity to actual IP recommendations. A global golden performance standard was defined and used along side the forward progress scores to supplement student-specific recommendations with appropriate event adjustments and guide continuous student progression throughout a training campaign.

The final model only requires the user, in this case an IP, to provide a student identification number and corresponding training event number for which they would like a flight plan recommendation. Recommendation generation, from the time user inputs are provided to the time an event set is output, only takes seconds provided PTN evaluation history data is readily available.

Industry leading artificial intelligence and applied statistics techniques were successfully implemented to devise a model that generates a set of graded flight events for an upcoming flight while discouraging overall progress stagnation. In it's prototypical state, this model is designed to be used as a tool to aid the IP's flight planning process rather than perform it fully.

VIII. Future Work

8.1 Overview

The inherent uniqueness of PTN and the AutoGradebook concept open a variety of focus areas for future research. Additional research on the flight event recommendation task provides an opportunity to further refine a model that effectively generates flight event recommendations. A variety of approaches can be taken in conducting future work for flight event recommendations. Recommended future research includes, but is not limited to, in-depth model parameter tuning, incorporation of progression rates into projected FPS scores, the inclusion of event occurrence correlations, creation of additional collaborative-based progress standards, and creation of a reinforcement learning mechanism to increase model autonomy and performance over time.

8.2 Content-Based Model Tuning and Alternatives

Minor parameter tuning was conducted in this study. Future related work may emphasize improvement on the content-based model through parameter tuning given the collection of data from more recent PTN classes. Model fitting parameters to consider include but are not limited to input sequence lengths, number of hidden layers, training duration, activation functions, number of epochs, and training batch sizes. Alternative model structures may also be explored for better recommendations. Temporal Convolutional Networks have been shown to capture temporal data trends for sequential predictions. Alternatively, sequential set generation by individual events may also provide a valid approach to the event recommendation task.

8.2.1 TCN Model

Recommending the order that events are performed in flight is outside of the scope of this research. Therefore, the order of event recommendation is irrelevant, as well. LSTM RNN models and take a series of vectors as input and output a single vector for the given task. The resulting vector suggests a specific order of events. Given that the order of the recommended events does not matter for the given task, a set-based output may be more suitable. Rather than generating an entire event set at once, events for the next flight may also be individually generated.

A TCN model structure is another approach at performing multi-output binary classification. TCNs function similarly to LSTM RNNs by using a series of evaluation data to predict the next-exercise event occurrences. However, a TCN uses the same series of data to individually predict the occurrence of events rather than predicting the set as a whole.

8.2.2 Sequential Set Generation Model

A model for sequential set generation would suggest more advanced machine learning techniques than a TCN model incorporating relationships between the recommended events while still selecting events individually. A model for sequential set generation may follow a similar design to automated text generation models that use individual characters of the alphabet to predict the next characters. Individual event generation provides the ability to build an even more personalized event set, accounting for student progress as well as event pairings. IBM researchers proposed a method for predicting set-valued outputs in the Proceedings of the AAAI Conference on Artificial Intelligence that may apply directly to the PTN event planning task [19].

8.3 Event Utility

Current event utilities are calculated independently. Including a component that accounts for the probability of events occurring simultaneously with events in the initial content-based recommendation may help provide more appropriate event swaps from the collaborative-based model. Calculating expected probabilities using bayesian statistics may be a feasible approach to this function.

8.4 Incorporation of Probability of Progress

The data shows that students are not guaranteed to receive a higher score each additional time they are evaluated on an event. Therefore, is it an broad generalization to assume that students will improve when predicting FPS for students performing above the golden performance standard. Incorporating a component to account for a student's probability of advancing to the next highest letter grade may improve FPS scores. The probability of advancing can be calculated as a collaborative statistic from the global dataset. Accounting for the probability of advancing rather than assuming advancement would result in a more accurate expected value for IEF in the next training exercise and thus result in a more accurate FPS prediction for students operating above the golden standard.

8.5 Additional Performance Standards for Tailored Recommendation

More tailored performance standards may provide more personalization for recommendations and more specific training guidance than a single global golden standard. Progressing toward a local standard is more logical for lower performing students to avoid over extreme, or forced, acceleration through a program. If students are only being pushed toward a global golden standard, lower performing students may continuously be introduced to more advanced events to catch up to the standard, despite receiving lower scores on the events. This results in skewed recommendations. Moving toward a local standard allows recommendations to gradually progress lower performing students through training without creating gaps in the learning process.

The golden standard becomes less extreme the higher a students performance is. Therefore, higher performing students may receive recommendations based on the golden standard that progress them through training quicker and with less risk of failure.

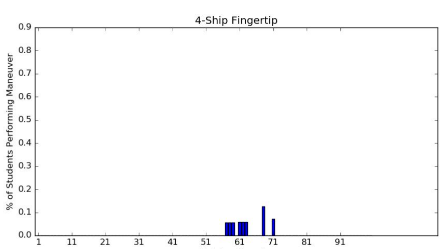
The proposed model does not apply multiple performance standards for recommendation guidance in order to make recommendations that consistently challenge student performance. Making recommendations towards a local standard of performance encourages extreme personalization to an individual's progress, which may result in slower advancement through training. Therefore, tailoring recommendations to more personalized performance standards based on a student's skill level at any given point throughout training does not align with the PTN program goals. However, incorporating this feature into the AutoGradebook could greatly benefit the individual training effectiveness of students in a less time-relevant environment. For example, moving towards more personalized performance standards may be an effective way to avoid learning gaps for AutoGradebook users pre-pilot training at service academies or universities.

Currently, the FPS score does not account for the possibility of student digression. The calculation incorporates all of a students training evaluations. However, to improve upon the FPS, a number of previous training exercises for review may be specified. Adding such a feature would update IEP scores and clarify if students required further training in a specific event rather than determining proficiency by the best overall evaluation score within their campaign.

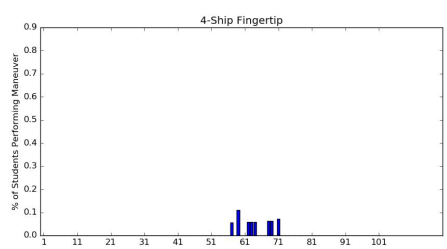
8.6 Reinforcement Learning Component

Adding a reinforcement learning component to the model can provide real-time model updates based on real instructor pilot feedback of recommendations. Real feedback is the best way to test and update recommender system models because of the inherent personalization of the recommendation process. However, this component can only be applied once PTN has developed a standardized data collection strategy.

Appendix A: Detailed Data Analysis Plots

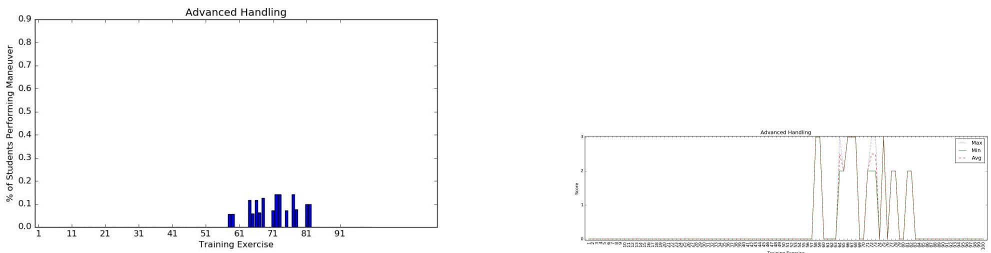


(a) 4-Ship Fingertip Occurrence



(b) 4-Ship Fingertip Evaluation Statistics

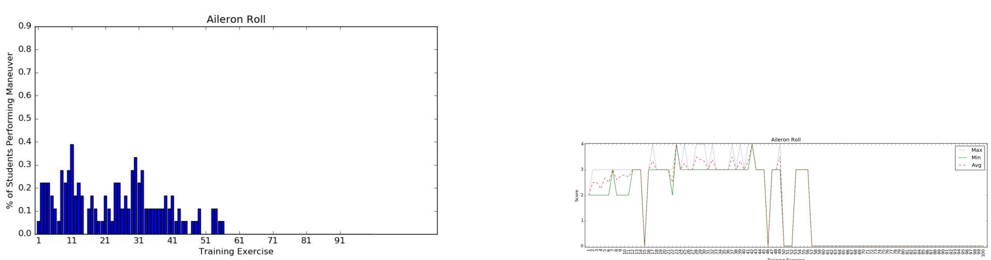
Figure 14. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for 4-Ship Fingertip



(a) Advanced Handling Occurrence

(b) Advanced Handling Statistics

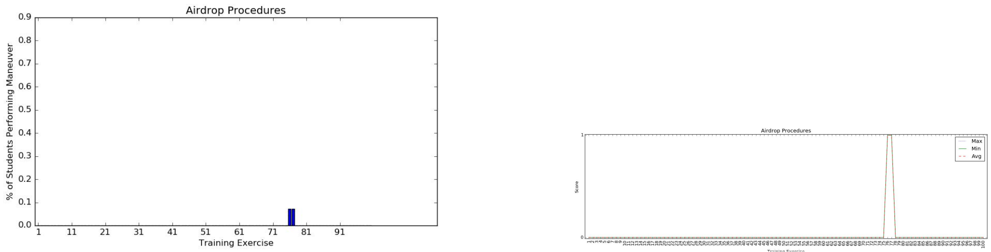
Figure 15. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Advanced Handling



(a) Aileron Roll Occurrence

(b) Aileron Roll Evaluation Statistics

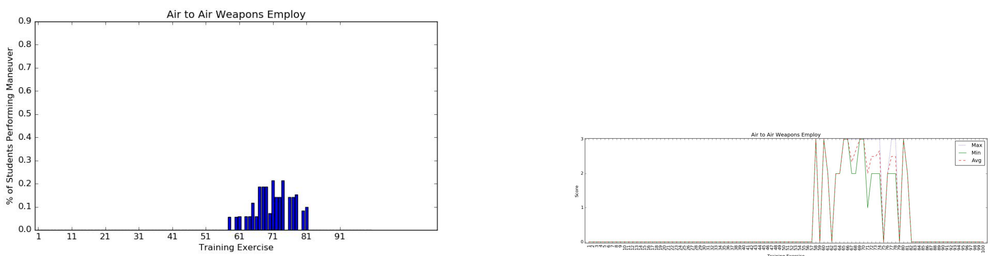
Figure 16. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Aileron Roll



(a) Airdrop Procedures Occurrence

(b) Airdrop Procedures Evaluation Statistics

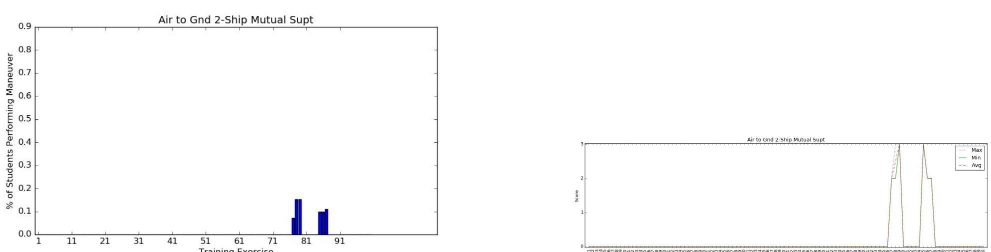
Figure 17. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Airdrop Procedures



(a) Air to Air Weapons Employ Occurrence

(b) Air to Air Weapons Employ Evaluation Statistics

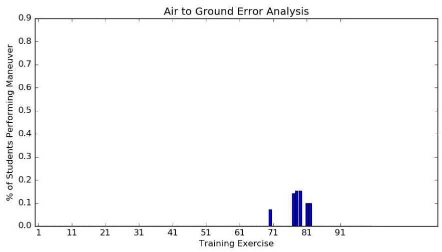
Figure 18. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Air to Air Weapons Employ



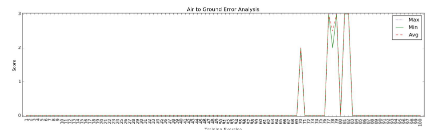
(a) Air to Gnd 2-Ship Mutual Supt Occurrence

(b) Air to Gnd 2-Ship Mutual Supt Statistics

Figure 19. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Air to Gnd 2-Ship Mutual Supt

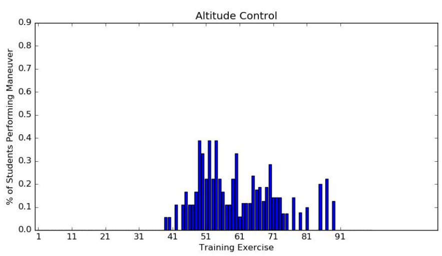


(a) Air to Ground Error Analysis Occurrence

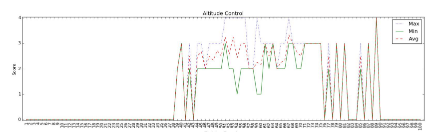


(b) Air to Ground Error Analysis Statistics

Figure 20. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Air to Ground Error Analysis

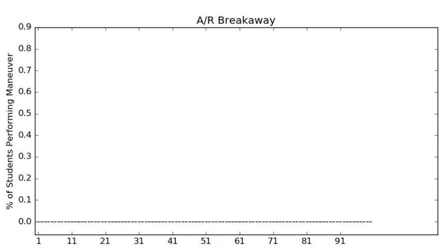


(a) Altitude Control Occurrence

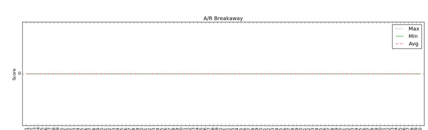


(b) Altitude Control Statistics

Figure 21. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Altitude Control

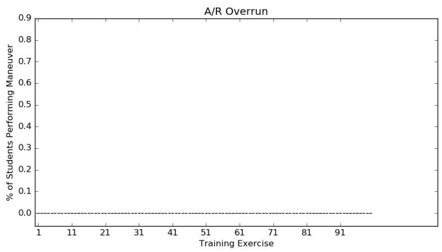


(a) A/R Breakaway Occurrence

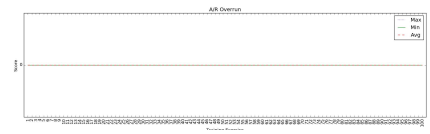


(b) A/R Breakaway Evaluation Statistics

Figure 22. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for A/R Breakaway

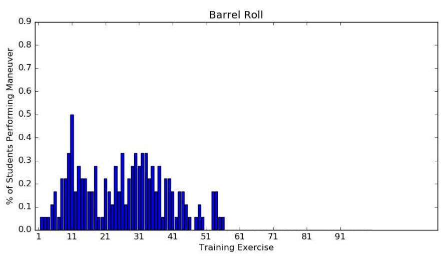


(a) A/R Overrun Occurrence

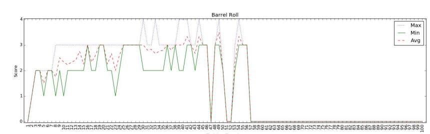


(b) A/R Overrun Statistics

Figure 23. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for A/R Overrun

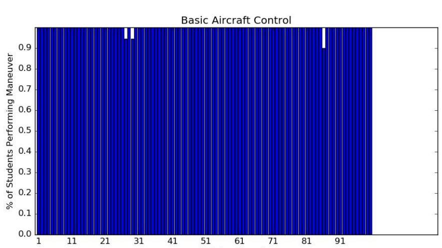


(a) Barrel Roll Occurrence

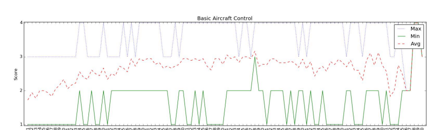


(b) Barrel Roll Evaluation Statistics

Figure 24. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Barrel Roll

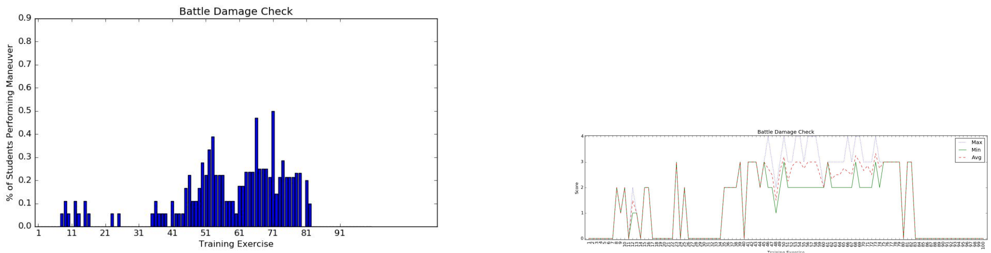


(a) Basic Aircraft Control Occurrence



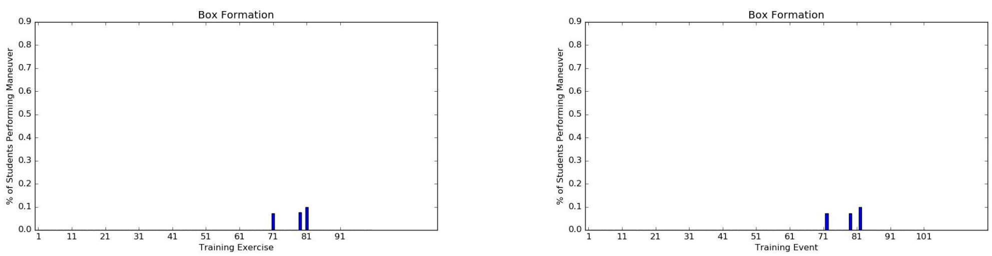
(b) Basic Aircraft Control Statistics

Figure 25. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Basic Aircraft Control



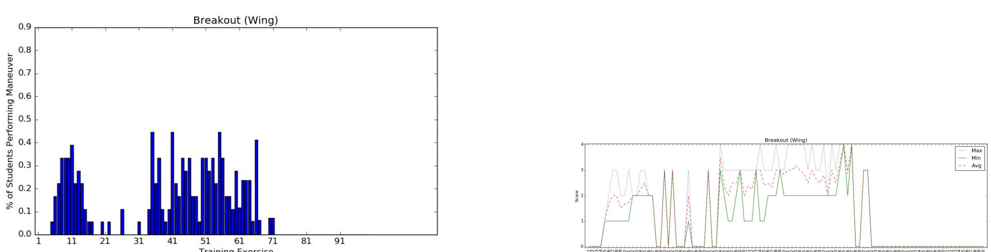
(a) Battle Damage Check Occurrence (b) Battle Damage Check Statistics

Figure 26. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Battle Damage Check



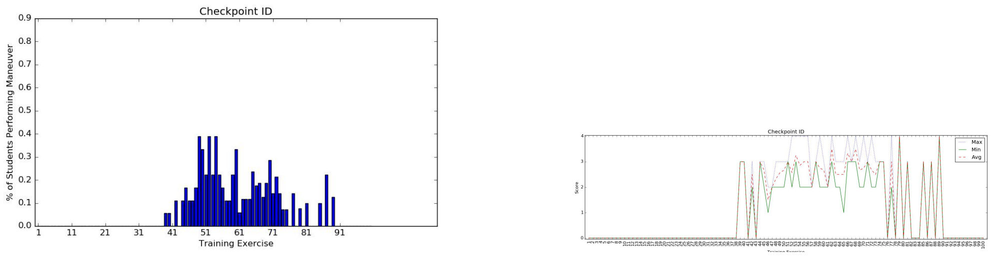
(a) Box Formation Occurrence (b) Box Formation Statistics

Figure 27. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Box Formation



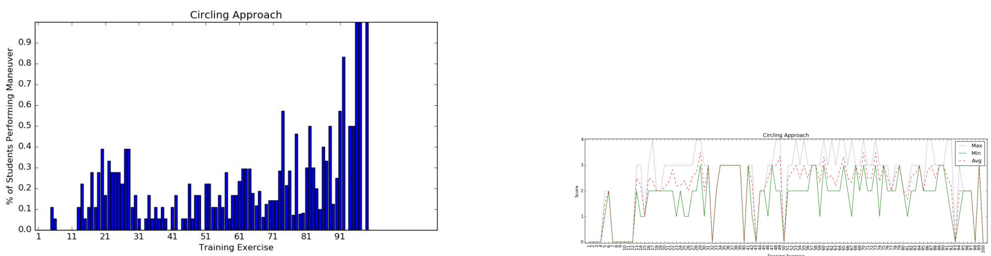
(a) Breakout (Wing) Occurrence (b) Breakout (Wing) Statistics

Figure 28. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Breakout (Wing)



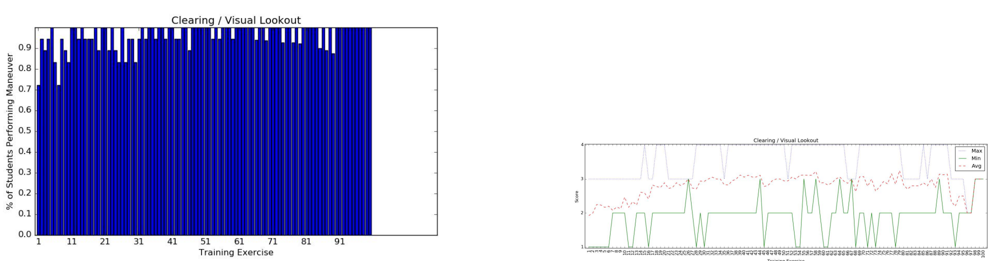
(a) Checkpoint ID Occurrence (b) Checkpoint ID Statistics

Figure 29. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Checkpoint ID



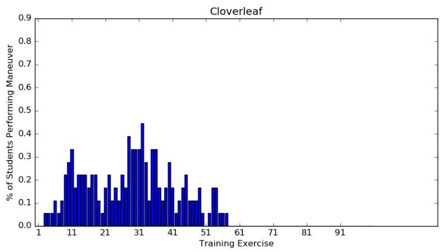
(a) Circling Approach Occurrence (b) Circling Approach Statistics

Figure 30. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Circling Approach

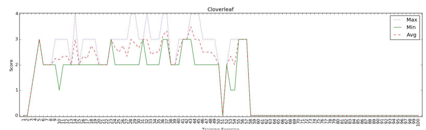


(a) Clearing/Visual Lookout Occurrence (b) Clearing/Visual Lookout Statistics

Figure 31. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Clearing/Visual Lookout

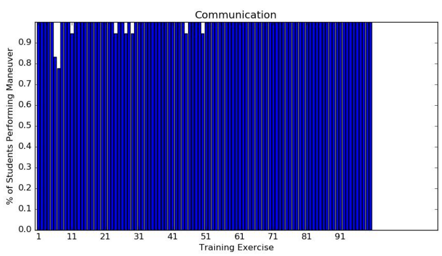


(a) Cloverleaf Occurrence

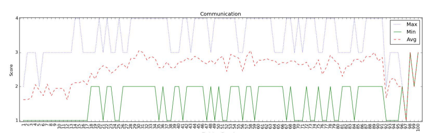


(b) Cloverleaf Statistics

Figure 32. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Cloverleaf

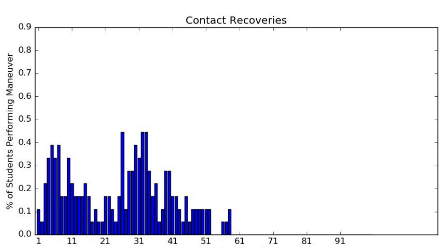


(a) Communication Occurrence

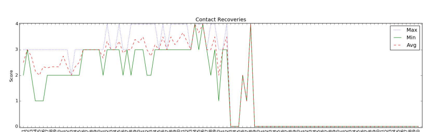


(b) Communication Statistics

Figure 33. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Communication

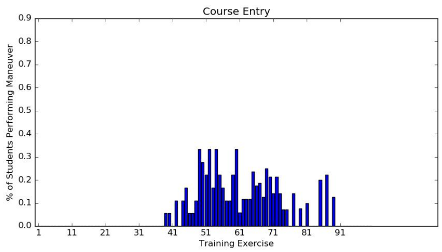


(a) Contact Recoveries Occurrence

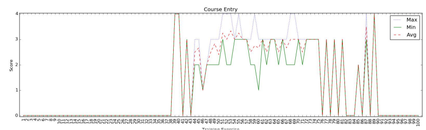


(b) Contact Recoveries Statistics

Figure 34. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Contact Recoveries

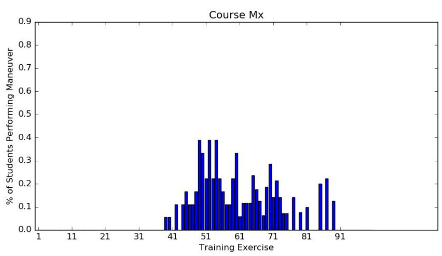


(a) Course Entry Occurrence

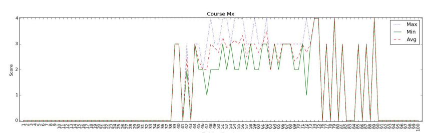


(b) Course Entry Statistics

Figure 35. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Course Entry

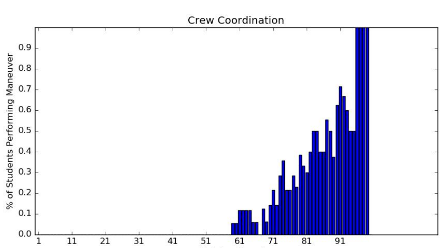


(a) Course Mx Occurrence

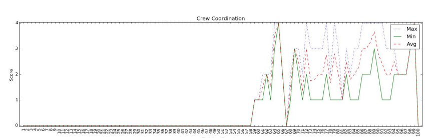


(b) Course Mx Statistics

Figure 36. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Course Mx

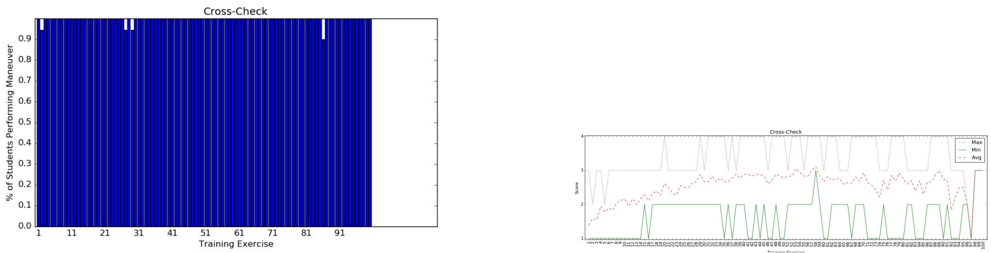


(a) Crew Coordination Occurrence



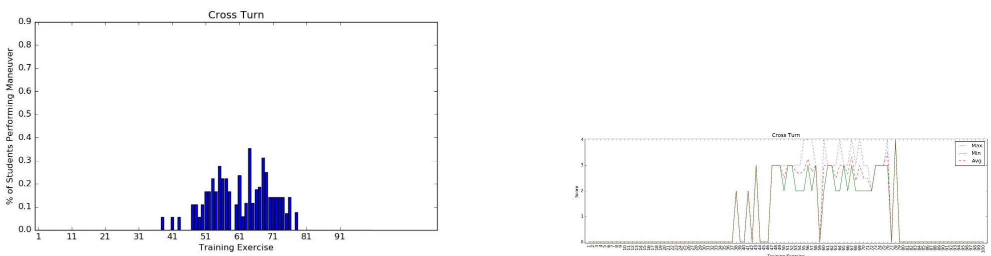
(b) Crew Coordination Statistics

Figure 37. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Crew Coordination



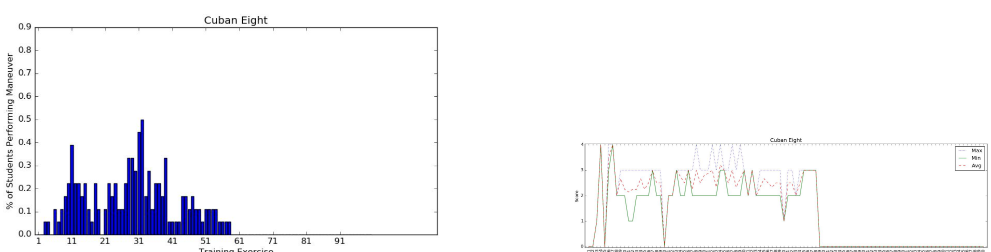
(a) Cross-Check Occurrence (b) Cross-Check Statistics

Figure 38. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Cross-Check



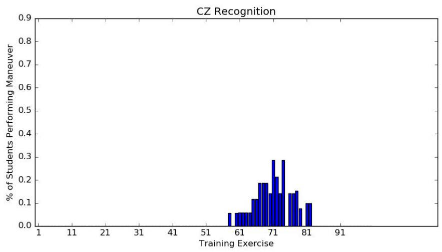
(a) Cross Turn Occurrence (b) Cross Turn Statistics

Figure 39. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Cross Turn

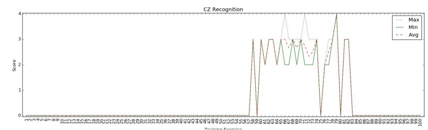


(a) Cuban Eight Occurrence (b) Cuban Eight Statistics

Figure 40. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Cuban Eight

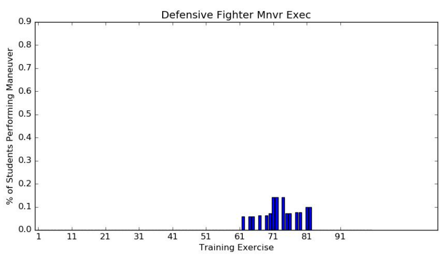


(a) CZ Recognition Occurrence

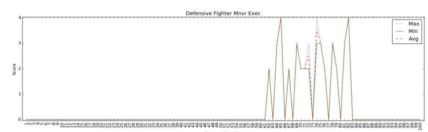


(b) CZ Recognition Statistics

Figure 41. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for CZ Recognition

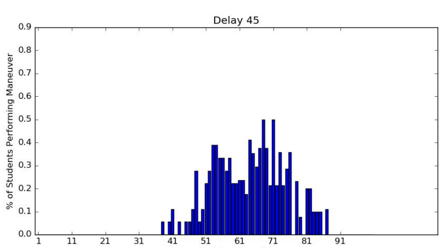


(a) Defensive Fighter Mnvr Exec Occurrence



(b) Defensive Fighter Mnvr Exec Statistics

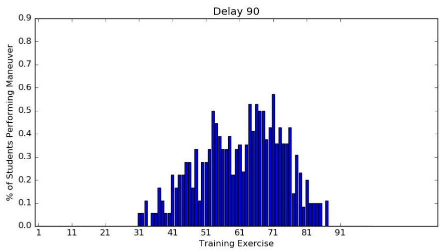
Figure 42. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Defensive Fighter Mnvr Exec



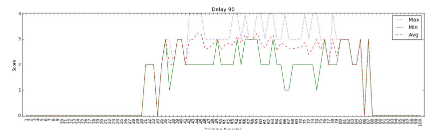
(a) Delay 45 Occurrence

(b) Delay 45 Statistics

Figure 43. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Delay 45

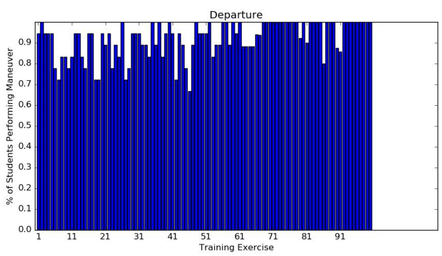


(a) Delay 90 Occurrence

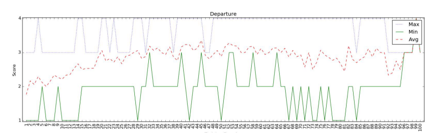


(b) Delay 90 Statistics

Figure 44. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Delay 90

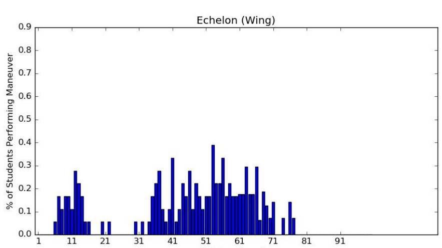


(a) Departure Occurrence

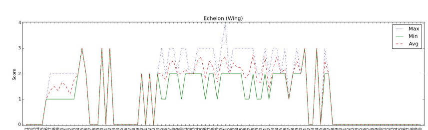


(b) Departure Statistics

Figure 45. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Departure

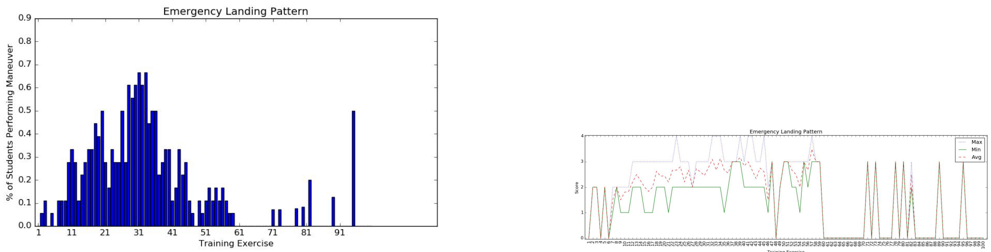


(a) Echelon (Wing) Occurrence



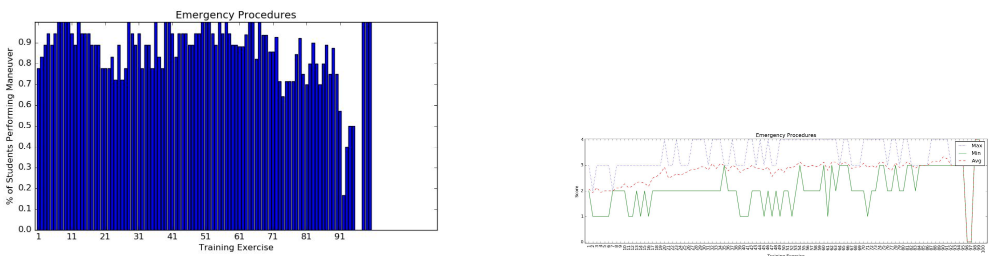
(b) Echelon (Wing) Statistics

Figure 46. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Echelon (Wing)



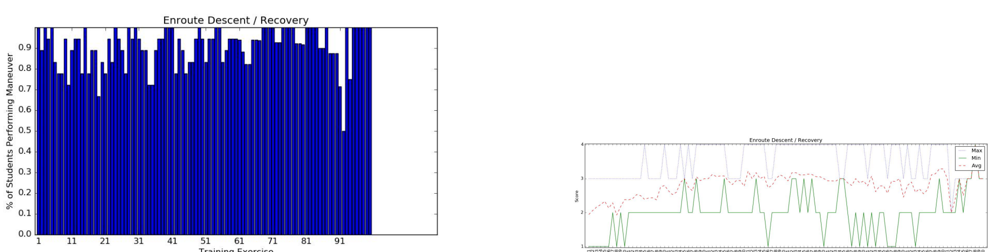
(a) Emergency Landing Pattern Occurrence (b) Emergency Landing Pattern Statistics

Figure 47. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Emergency Landing Pattern



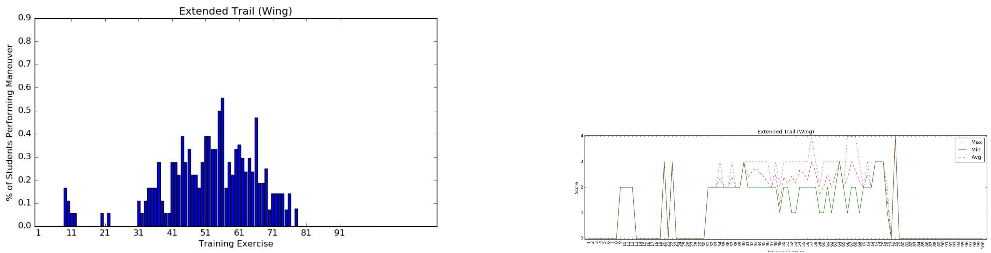
(a) Emergency Procedures Occurrence (b) Emergency Procedures Statistics

Figure 48. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Emergency Procedures



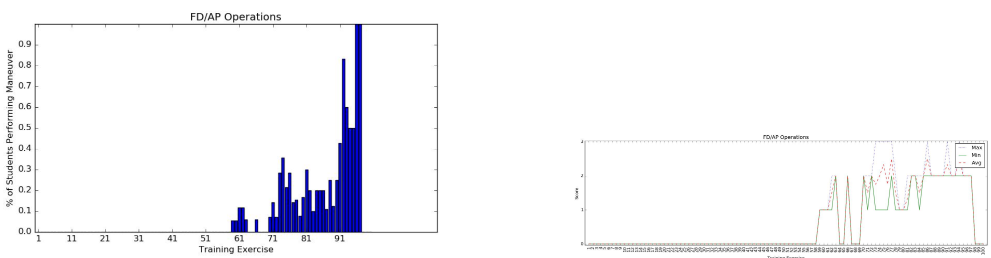
(a) Enroute Descent/Recovery Occurrence (b) Enroute Descent/Recovery Statistics

Figure 49. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Enroute Descent/Recovery



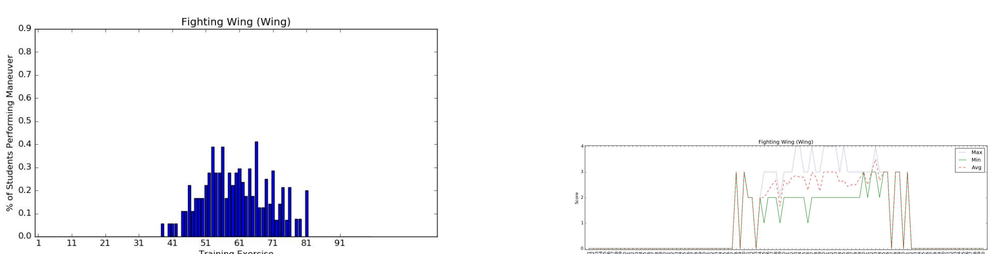
(a) Extended Trail (Wing) Occurrence (b) Extended Trail (Wing) Statistics

Figure 50. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Extended Trail (Wing)



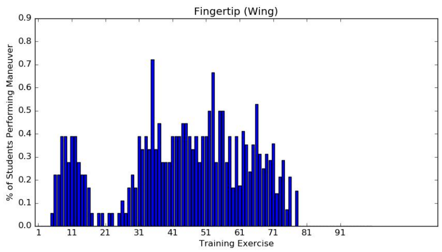
(a) FD/AP Operations Occurrence (b) FD/AP Operations Statistics

Figure 51. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for FD/AP Operations

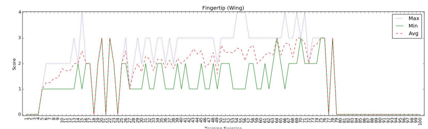


(a) Fighting Wing (Wing) Occurrence (b) Fighting Wing (Wing) Statistics

Figure 52. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Fighting Wing (Wing)

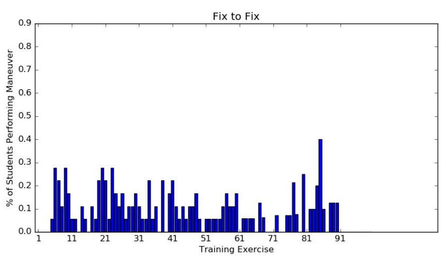


(a) Fingertip (Wing) Occurrence

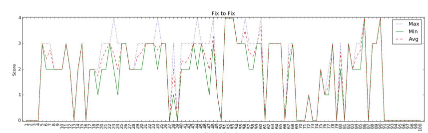


(b) Fingertip (Wing) Statistics

Figure 53. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Fingertip (Wing)

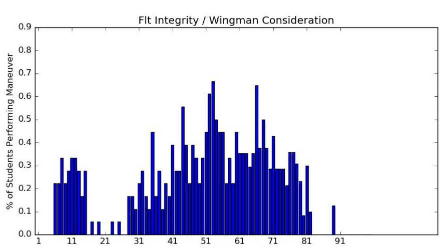


(a) Fix to Fix Occurrence

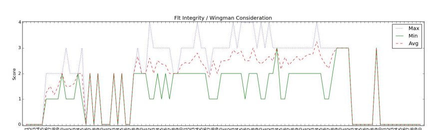


(b) Fix to Fix Statistics

Figure 54. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Fix to Fix

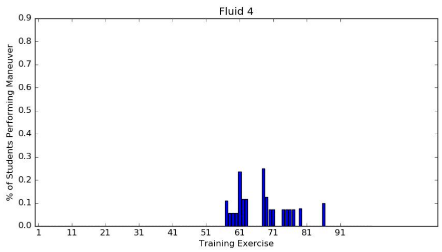


(a) Flt Integrity/Wingman Consideration Occurrence

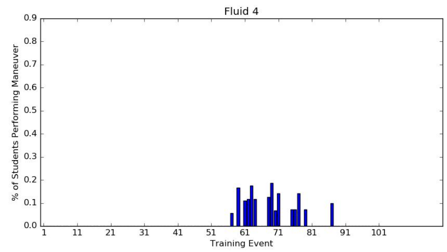


(b) Flt Integrity/Wingman Consideration Statistics

Figure 55. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Flt Integrity/Wingman Consideration

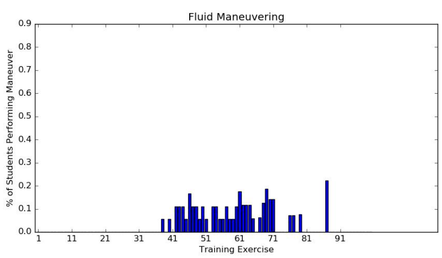


(a) Fluid4 Occurrence

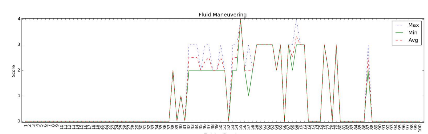


(b) Fluid4 Statistics

Figure 56. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Fluid4

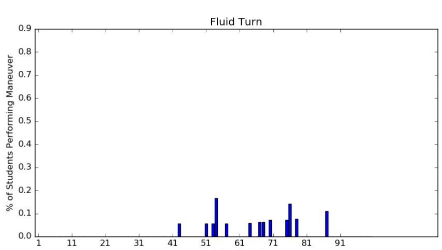


(a) Fluid Maneuvering Occurrence

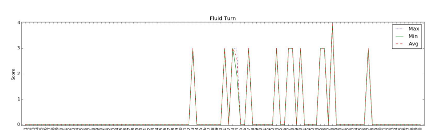


(b) Fluid Maneuvering Statistics

Figure 57. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Fluid Maneuvering

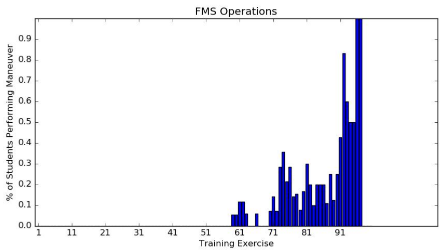


(a) Fluid Turn Occurrence

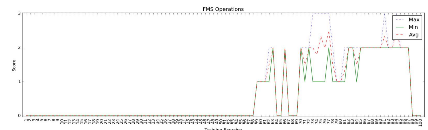


(b) Fluid Turn Statistics

Figure 58. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Fluid Turn

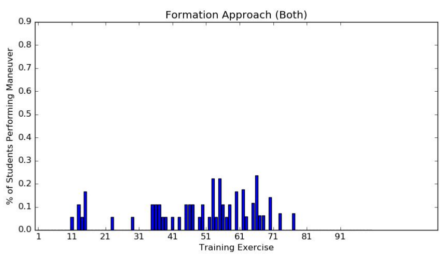


(a) FMS Operations Occurrence

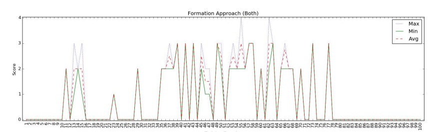


(b) FMS Operations Statistics

Figure 59. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for FMS Operations

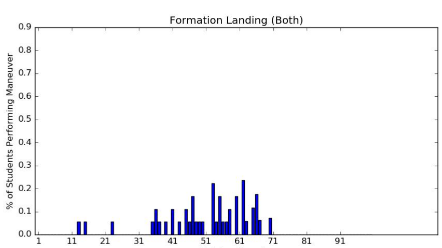


(a) Formation Approach (Both) Occurrence

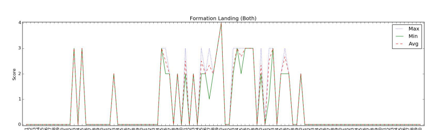


(b) Formation Approach (Both) Statistics

Figure 60. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Formation Approach (Both)

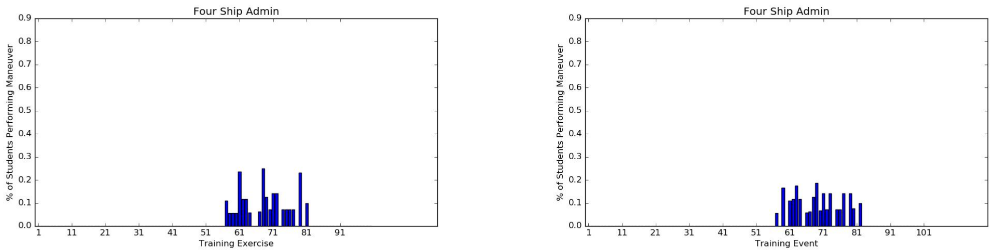


(a) Formation Landing (Both) Occurrence



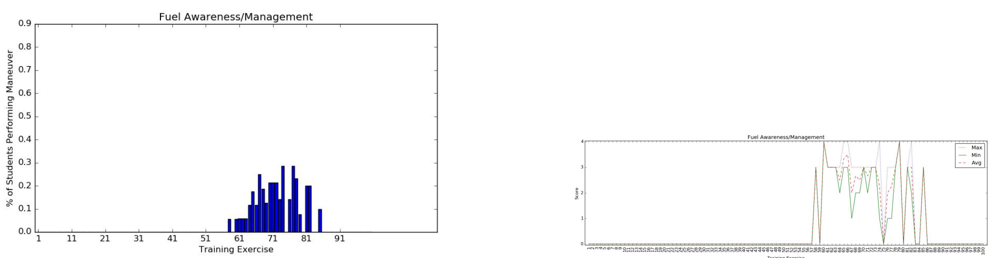
(b) Formation Landing (Both) Statistics

Figure 61. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Formation Landing (Both)



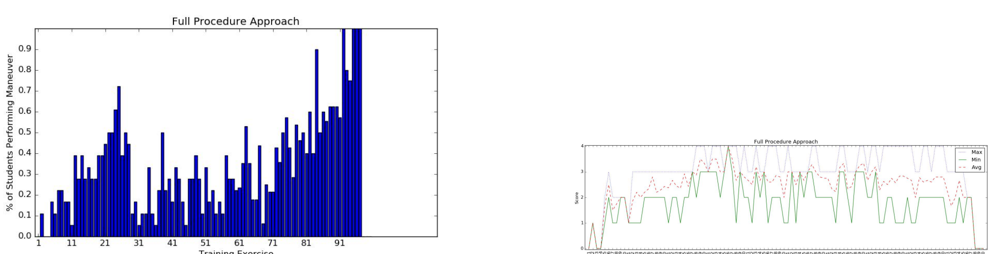
(a) Four Ship Admin Occurrence (b) Four Ship Admin Statistics

Figure 62. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Four Ship Admin



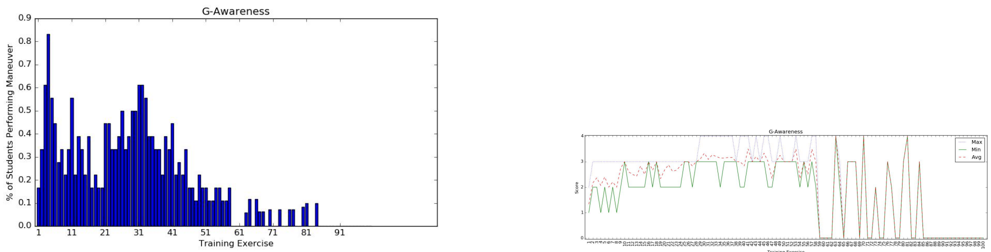
(a) Fuel Awareness/Management Occurrence (b) Fuel Awareness/Management Statistics

Figure 63. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Fuel Awareness/Management



(a) Full Procedure Approach Occurrence (b) Full Procedure Approach Statistics

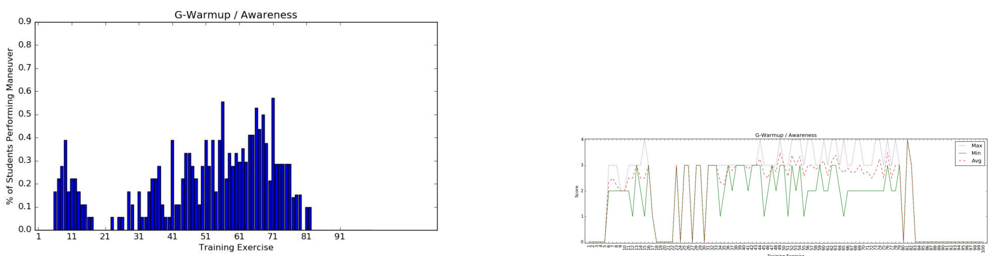
Figure 64. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Full Procedure Approach



(a) G-Awareness Occurrence

(b) G-Awareness Statistics

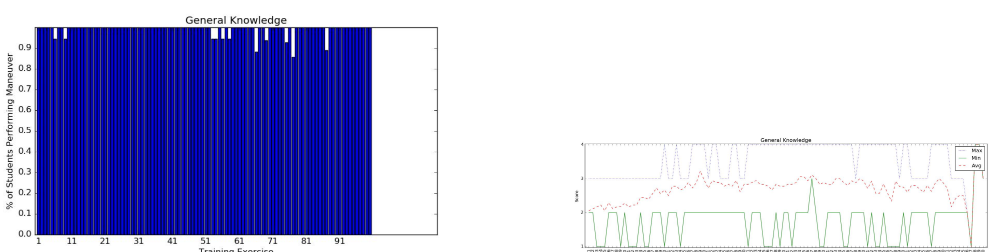
Figure 65. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for G-Awareness



(a) G-Warmup/Awareness Occurrence

(b) G-Warmup/Awareness Statistics

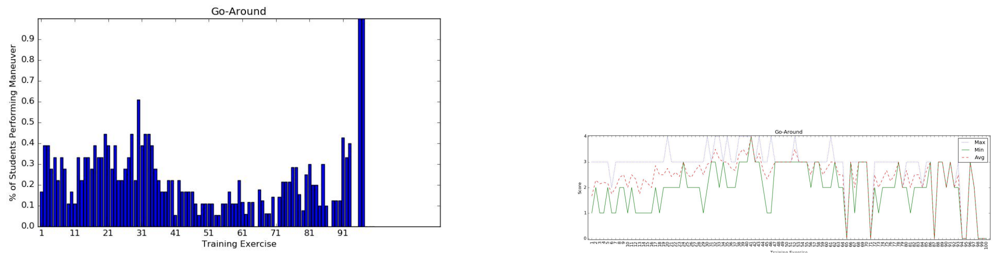
Figure 66. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for G-Warmup/Awareness



(a) General Knowledge Occurrence

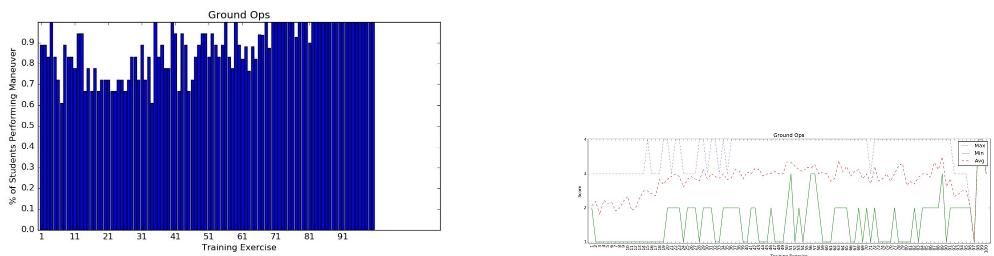
(b) General Knowledge Statistics

Figure 67. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for General Knowledge



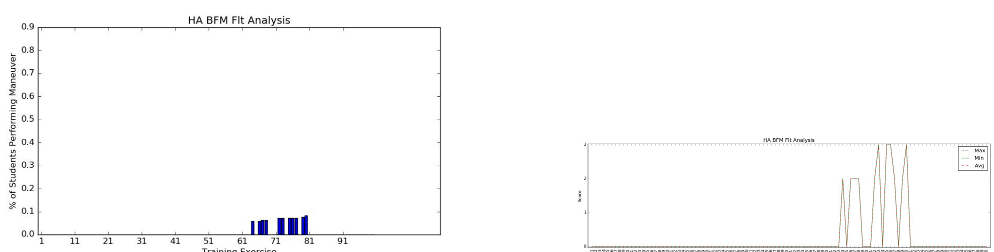
(a) Go-Around Occurrence (b) Go-Around Statistics

Figure 68. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Go-Around



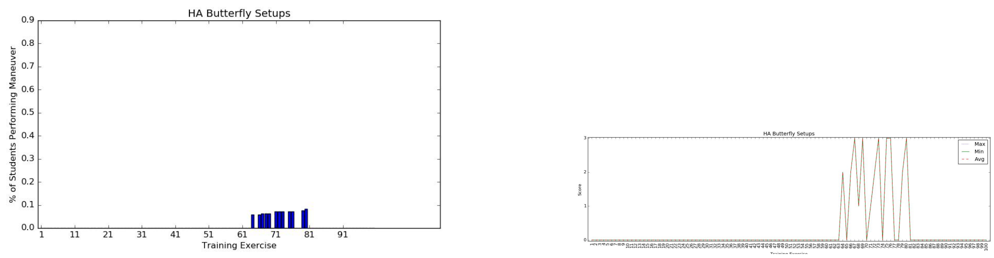
(a) Ground Ops Occurrence (b) Ground Ops Statistics

Figure 69. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Ground Ops



(a) HA BFM Flt Analysis Occurrence (b) HA BFM Flt Analysis Statistics

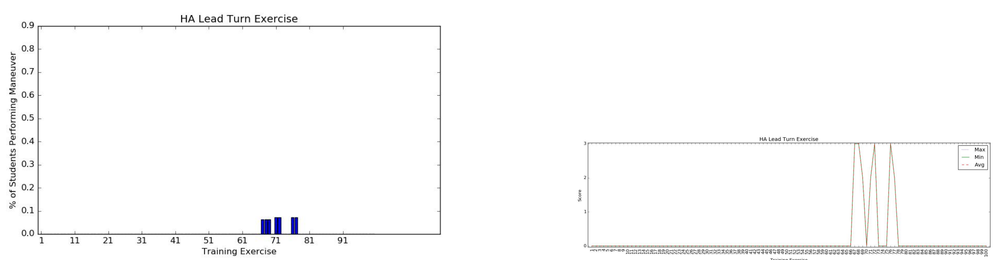
Figure 70. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for HA BFM Flt Analysis



(a) HA Butterfly Setups Occurrence

(b) HA Butterfly Setups Statistics

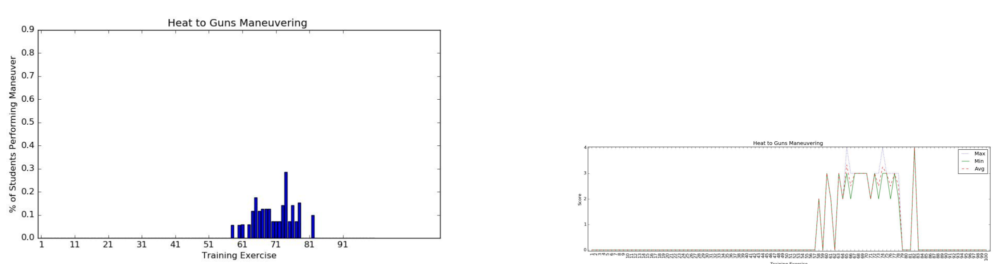
Figure 71. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for HA Butterfly Setups



(a) HA Lead Turn Exercise Occurrence

(b) HA Lead Turn Exercise Statistics

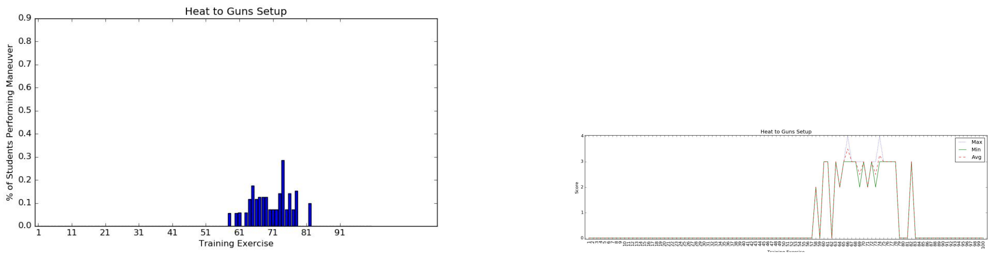
Figure 72. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for HA Lead Turn Exercise



(a) Heat to Guns Maneuvering Occurrence

(b) Heat to Guns Maneuvering Statistics

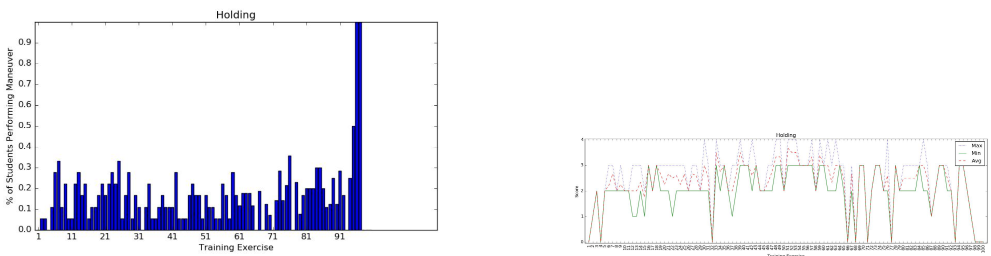
Figure 73. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Heat to Guns Maneuvering



(a) Heat to Guns Setup Occurrence

(b) Heat to Guns Setup Statistics

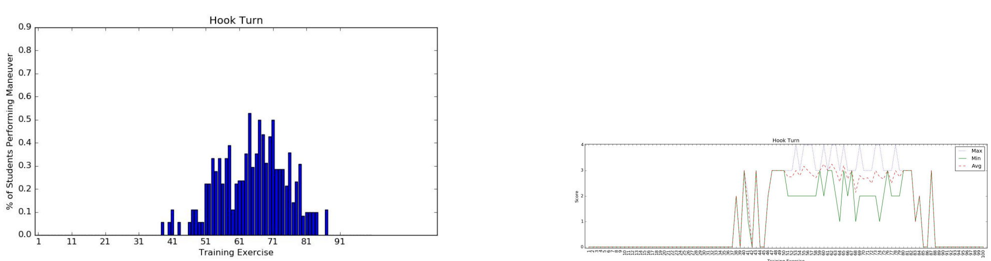
Figure 74. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Heat to Guns Setup



(a) Holding Occurrence

(b) Holding Statistics

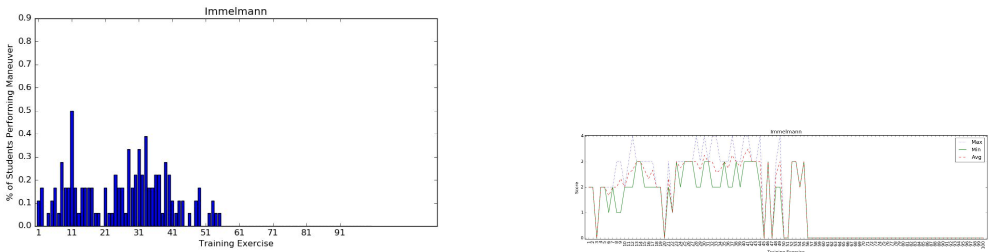
Figure 75. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Holding



(a) Hook Turn Occurrence

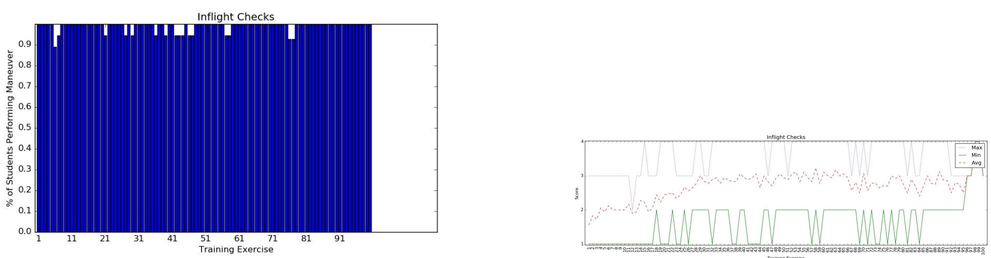
(b) Hook Turn Statistics

Figure 76. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Hook Turn



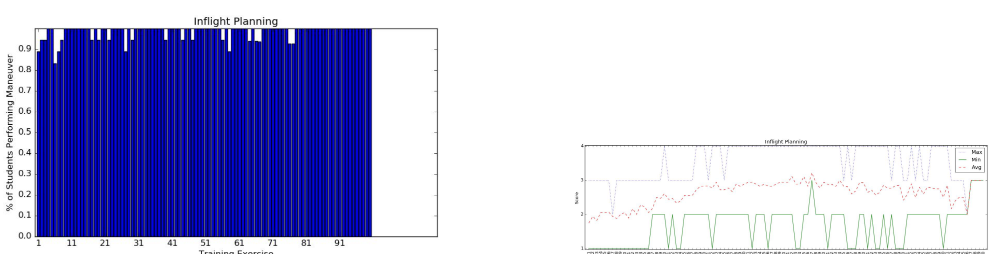
(a) Immelmann Occurrence (b) Immelmann Statistics

Figure 77. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Immelmann



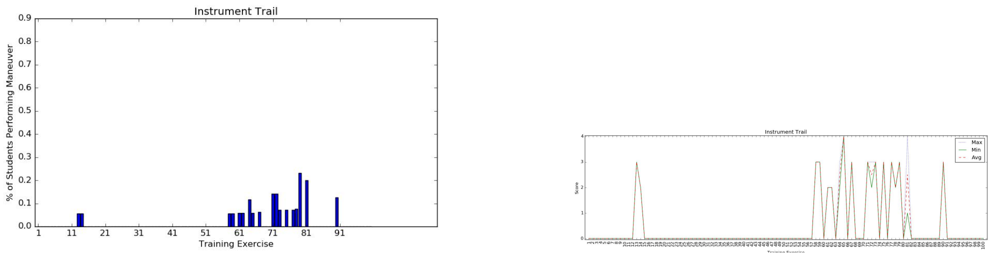
(a) Inflight Checks Occurrence (b) Inflight Checks Statistics

Figure 78. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Inflight Checks



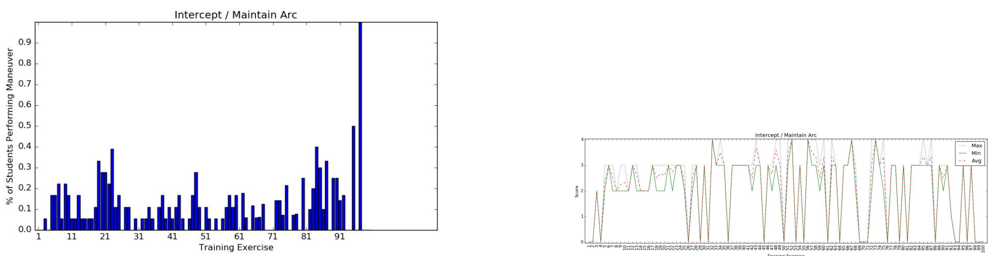
(a) Inflight Planning Occurrence (b) Inflight Planning Statistics

Figure 79. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Inflight Planning



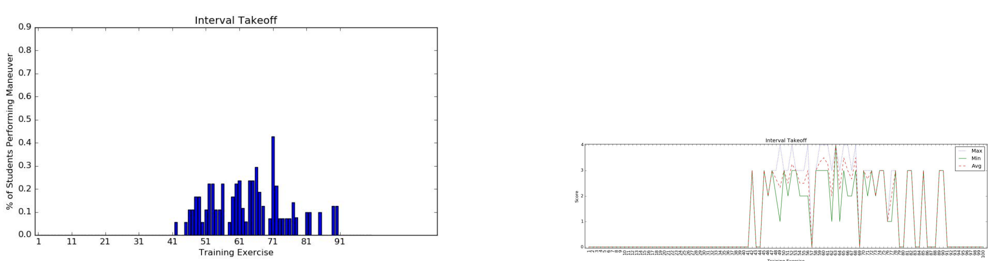
(a) Instrument Trail Occurrence (b) Instrument Trail Statistics

Figure 80. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Instrument Trail



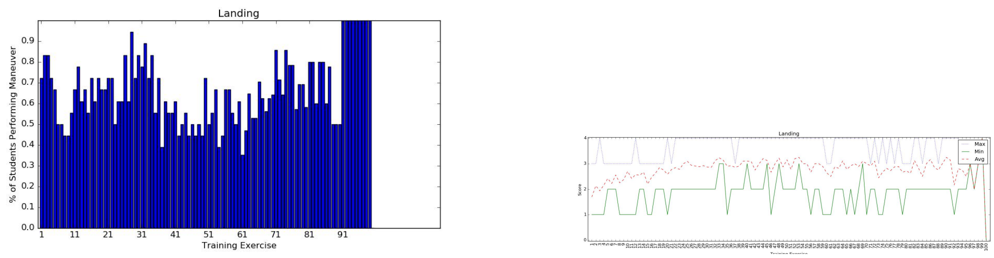
(a) Intercept/Maintain Arc Occurrence (b) Intercept/Maintain Arc Statistics

Figure 81. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Intercept/Maintain Arc



(a) Interval Takeoff Occurrence (b) Interval Takeoff Statistics

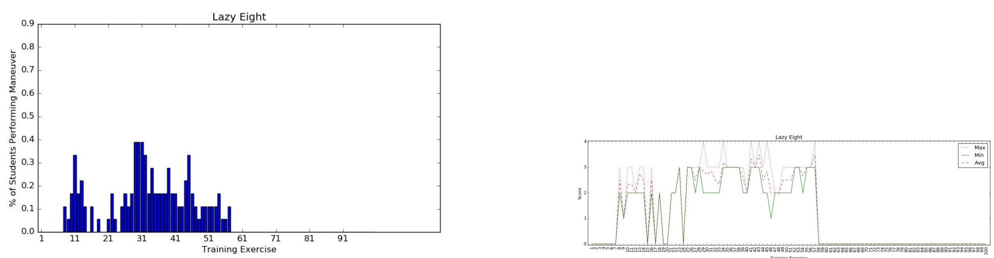
Figure 82. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Interval Takeoff



(a) Landing Occurrence

(b) Landing Statistics

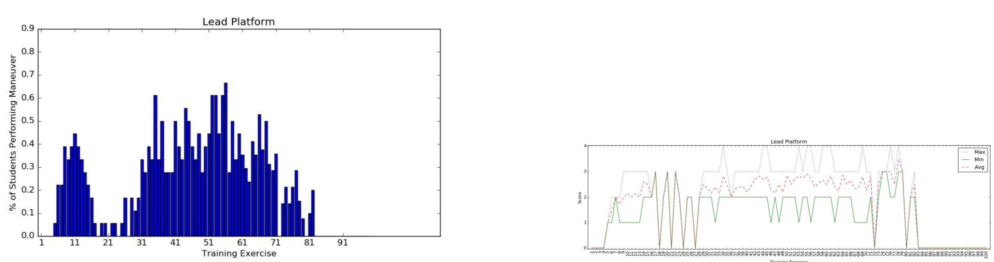
Figure 83. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Landing



(a) Lazy Eight Occurrence

(b) Lazy Eight Statistics

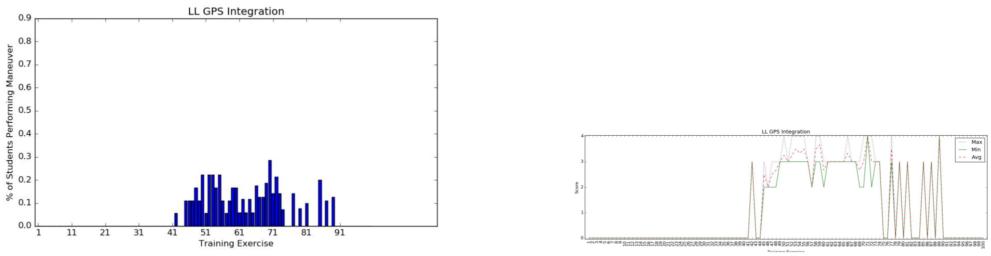
Figure 84. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Lazy Eight



(a) Lead Platform Occurrence

(b) Lead Platform Statistics

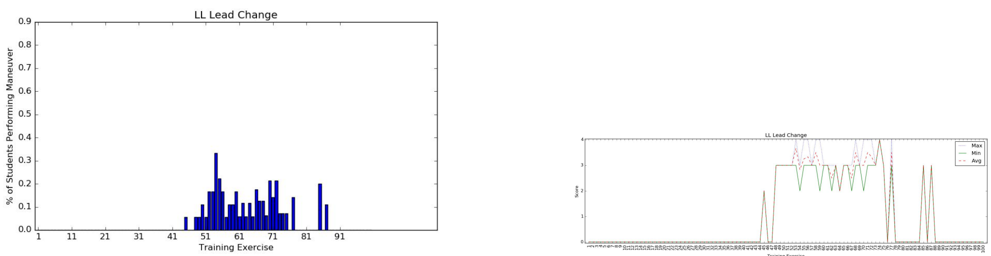
Figure 85. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Lead Platform



(a) LL GPS Integration Occurrence

(b) LL GPS Integration Statistics

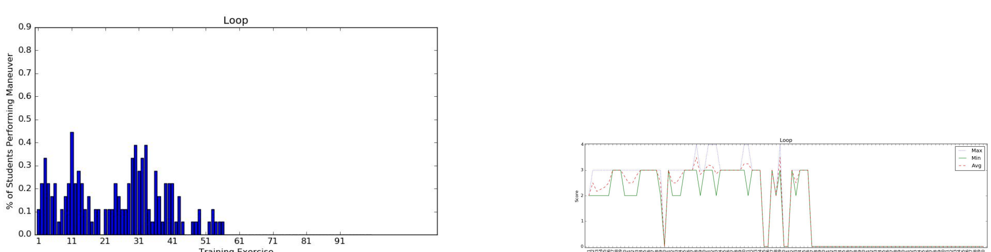
Figure 86. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for LL GPS Integration



(a) LL Lead Change Occurrence

(b) LL Lead Change Statistics

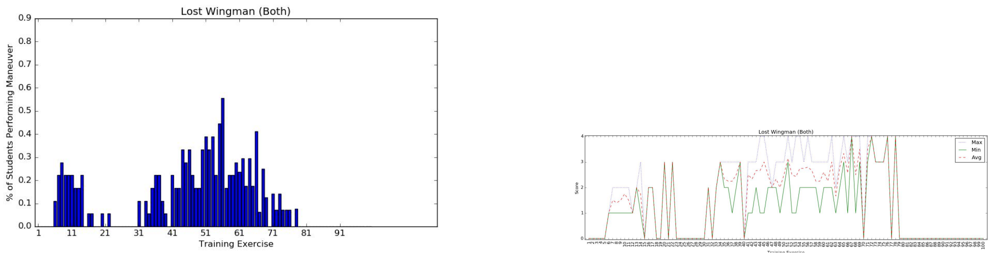
Figure 87. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for LL Lead Change



(a) Loop Occurrence

(b) Loop Evaluation Statistics

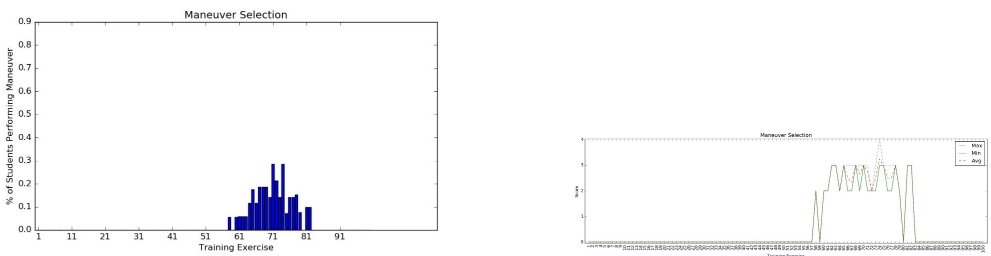
Figure 88. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Loop



(a) Lost Wingman (Both) Occurrence

(b) Lost Wingman (Both) Statistics

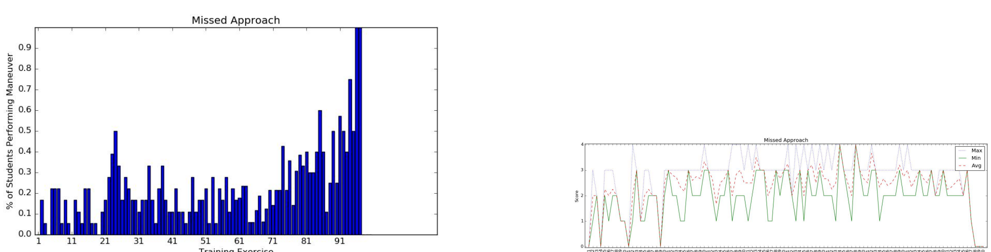
Figure 89. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Lost Wingman (Both)



(a) Maneuver Selection Occurrence

(b) Maneuver Selection Statistics

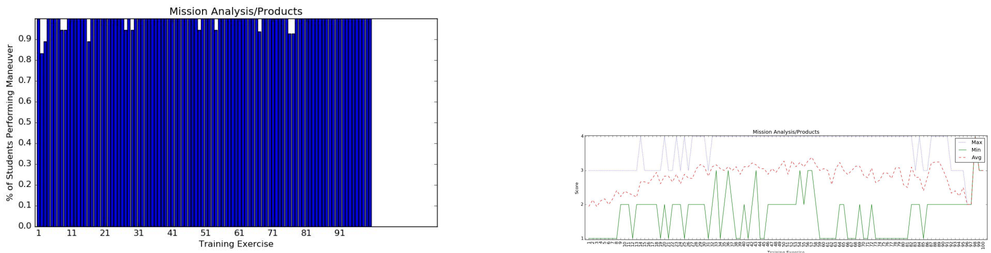
Figure 90. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Maneuver Selection



(a) Missed Approach Occurrence

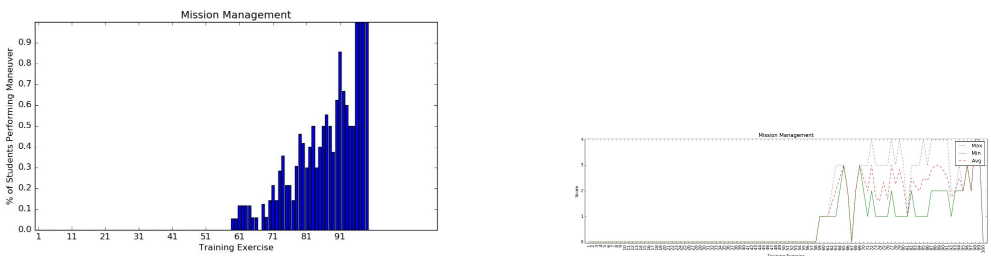
(b) Missed Approach Statistics

Figure 91. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Missed Approach



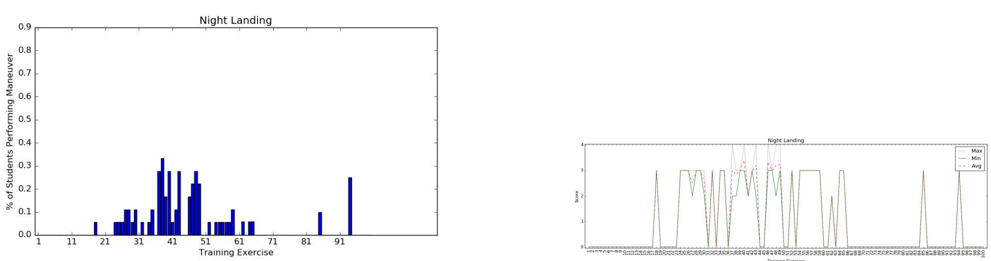
(a) Mission Analysis/Products Occurrence (b) Mission Analysis/Products Statistics

Figure 92. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Mission Analysis/Products



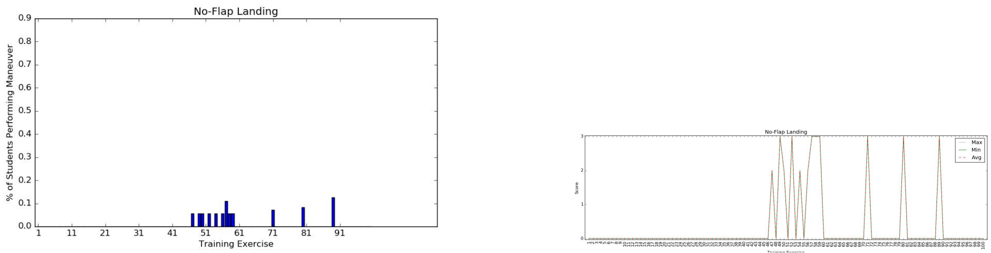
(a) Mission Management Occurrence (b) Mission Management Statistics

Figure 93. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Mission Management



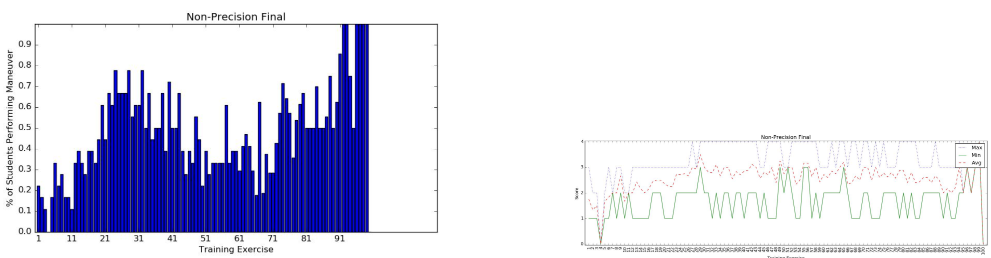
(a) Night Landing Occurrence (b) Night Landing Statistics

Figure 94. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Night Landing



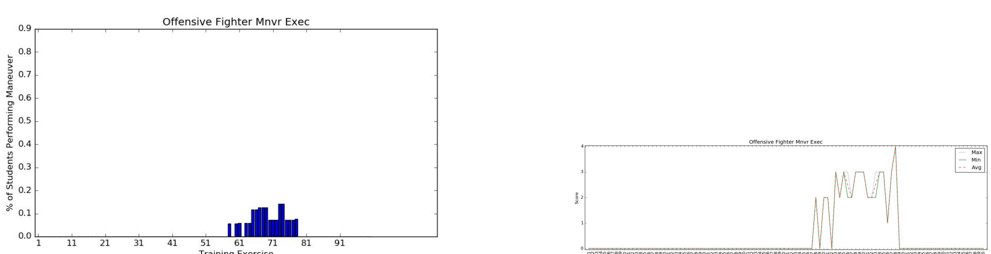
(a) No-Flap Landing Occurrence (b) No-Flap Landing Statistics

Figure 95. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for No-Flap Landing



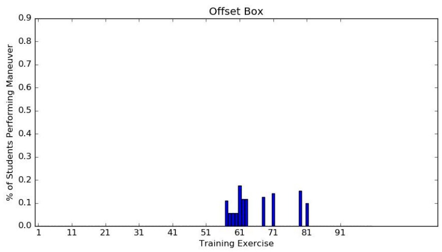
(a) Non-Precision Final Occurrence (b) Non-Precision Final Statistics

Figure 96. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Non-PrecisionFinal

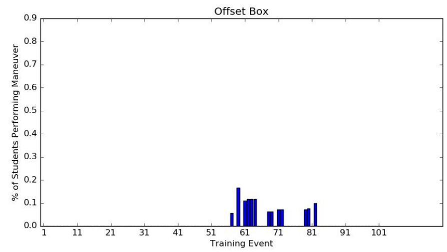


(a) Offensive Fighter Mnvr Exec Occurrence (b) Offensive Fighter Mnvr Exec Statistics

Figure 97. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Offensive Fighter Mnvr Exec

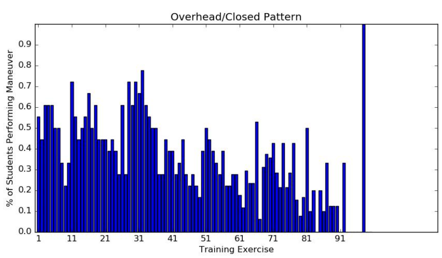


(a) Offset Box Occurrence

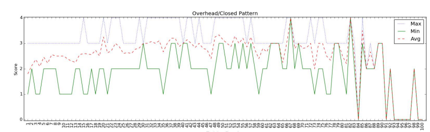


(b) Offset Box Statistics

Figure 98. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Offset Box

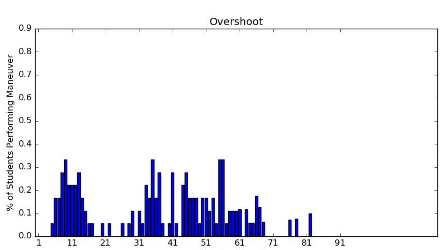


(a) Overhead/Closed Pattern Occurrence

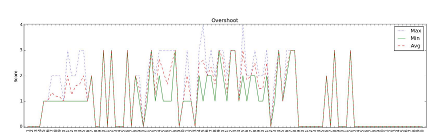


(b) Overhead/Closed Pattern Statistics

Figure 99. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Overhead Closed/Pattern

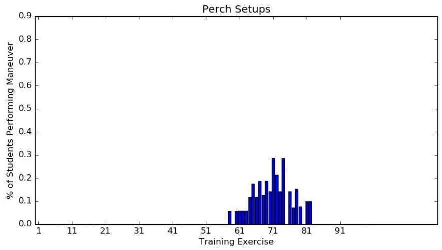


(a) Overshoot Occurrence

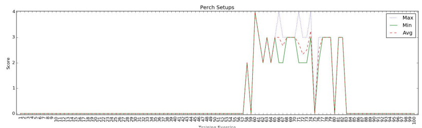


(b) Overshoot Statistics

Figure 100. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Overshoot

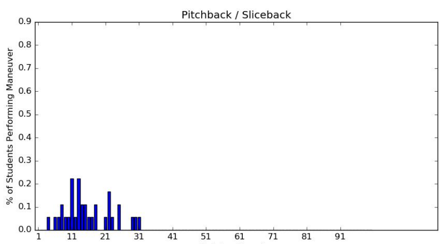


(a) Perch Setups Occurrence

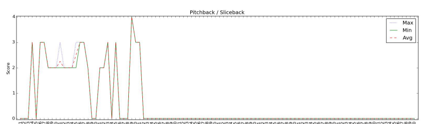


(b) Perch Setups Statistics

Figure 101. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Perch Setups

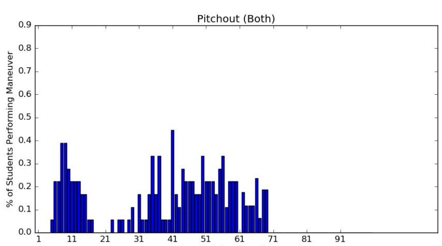


(a) Pitchback/Sliceback Occurrence

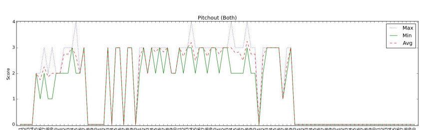


(b) Pitchback/Sliceback Statistics

Figure 102. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Pitchback/Sliceback

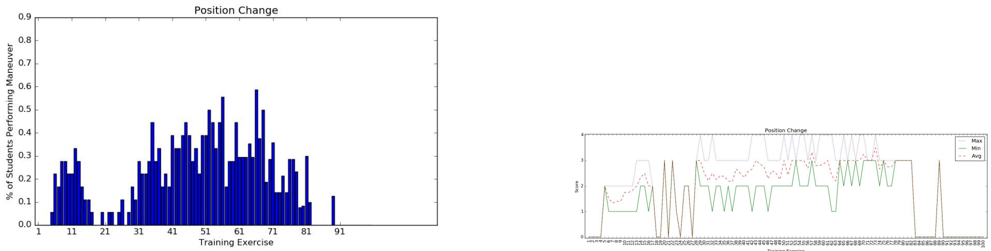


(a) Pitchout (Both) Occurrence



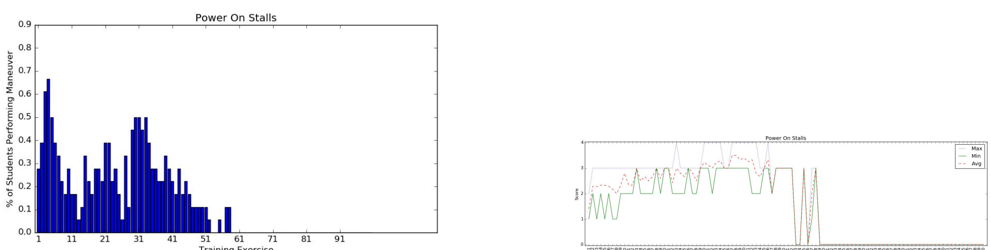
(b) Pitchout (Both) Statistics

Figure 103. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Pitchout (Both)



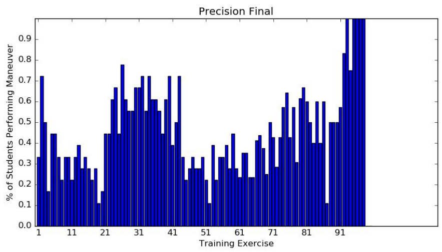
(a) Position Change Occurrence (b) Position Change Statistics

Figure 104. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Position Change

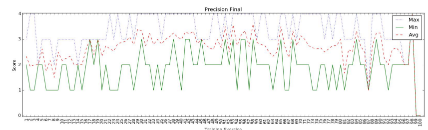


(a) Power On Stalls Occurrence (b) Power On Stalls Statistics

Figure 105. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Power On Stalls

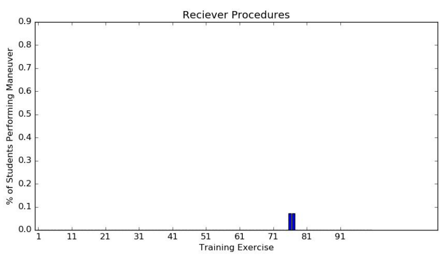


(a) Precision Final Occurrence

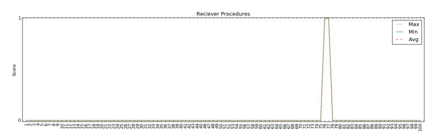


(b) Precision Final Statistics

Figure 106. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Precision Final

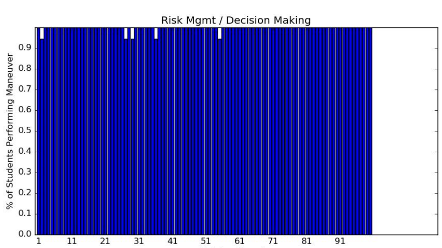


(a) Receiver Procedures Occurrence

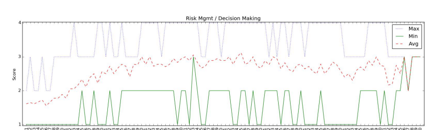


(b) Receiver Procedures Statistics

Figure 107. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Receiver Procedures

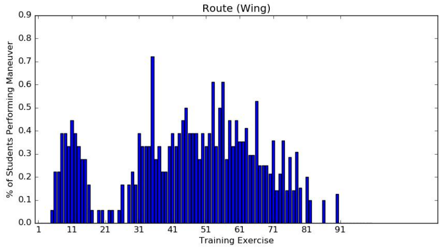


(a) Risk Mgmt/Decision Making Occurrence

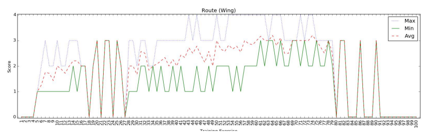


(b) Risk Mgmt/Decision Making Statistics

Figure 108. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Risk Mgmt/Decision Making

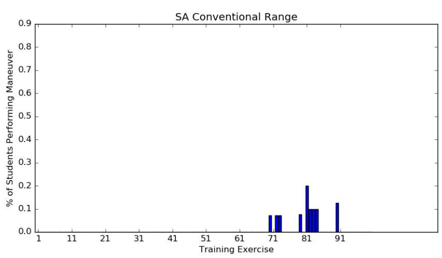


(a) Route (Wing) Occurrence

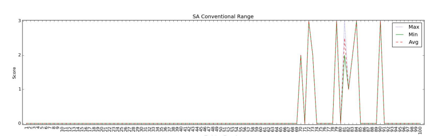


Statistics]
(b)
Route
(Wing)

Figure 109. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Route (Wing)

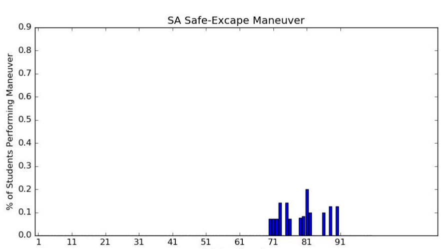


(a) SA Conventional Range Occurrence

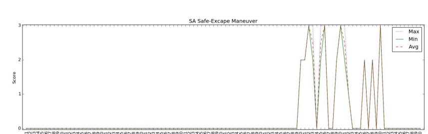


(b) SA Conventional Range Statistics

Figure 110. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for SA Conventional Range

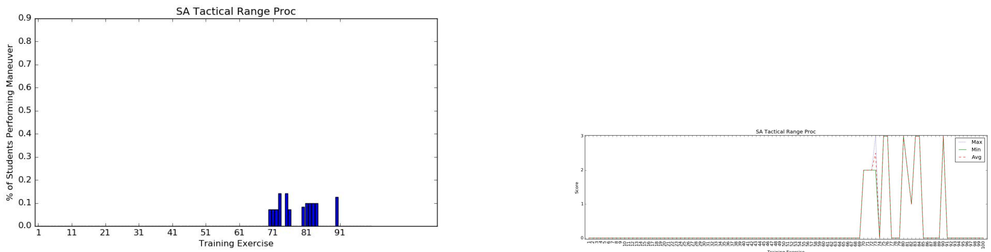


(a) SA Safe-Excape Maneuver Occurrence



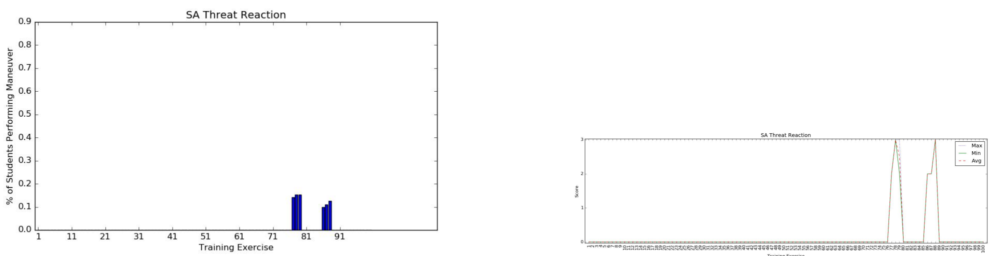
(b) SA Safe-Excape Maneuver Statistics

Figure 111. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for all SA Safe-Excape Maneuver



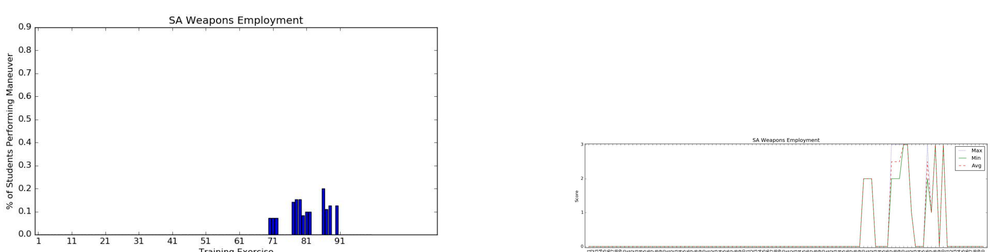
(a) SA Tactical Range Proc Occurrence (b) SA Tactical Range Proc Statistics

Figure 112. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for SA Tactical Range Proc



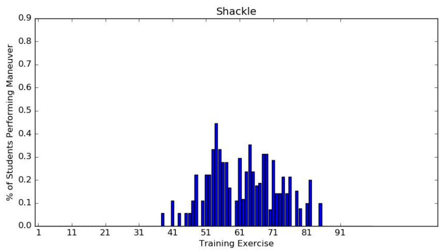
(a) SA Threat Reaction Occurrence (b) SA Threat Reaction Statistics

Figure 113. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for SA Threat Reaction

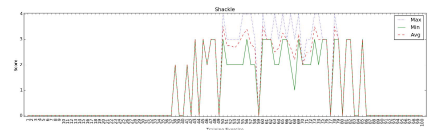


(a) SA Weapons Employment Occurrence (b) SA Weapons Employment Statistics

Figure 114. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for SA Weapons Employment

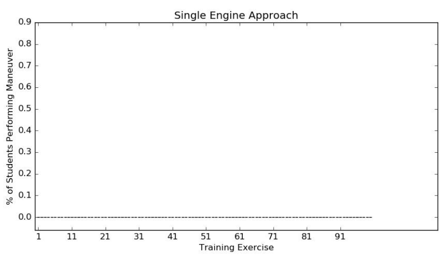


(a) Shackle Occurrence

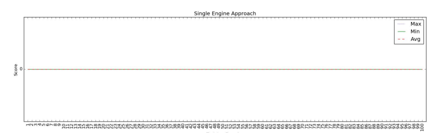


(b) Shackle Statistics

Figure 115. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Shackle

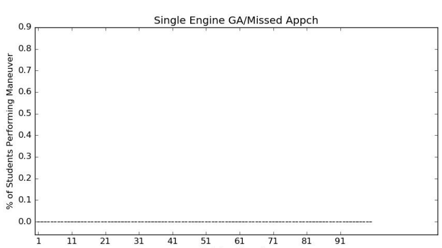


(a) Single Engine Approach Occurrence

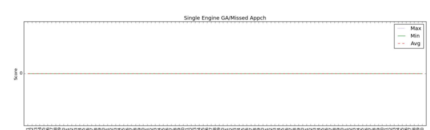


(b) Single Engine Approach Statistics

Figure 116. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Single Engine Approach

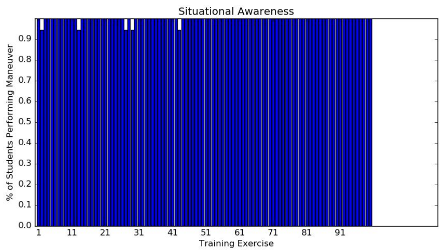


(a) Single Engine GA/Missed Appch Occurrence

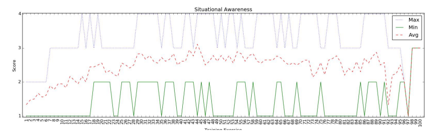


(b) Single Engine GA/Missed Appch Statistics

Figure 117. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Single Engine GA/Missed Appch

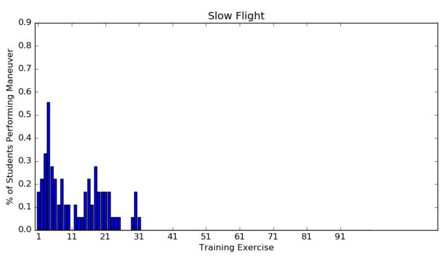


(a) Situational Awareness Occurrence

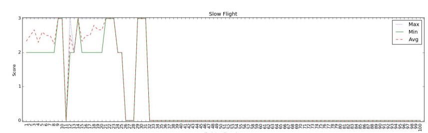


(b) Situational Awareness Statistics

Figure 118. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Situational Awareness

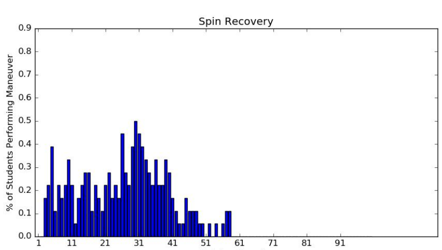


(a) Slow Flight Occurrence

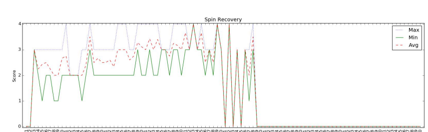


(b) Slow Flight Statistics

Figure 119. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Slow Flight

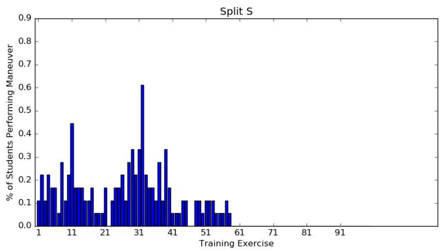


(a) Spin Recovery Occurrence

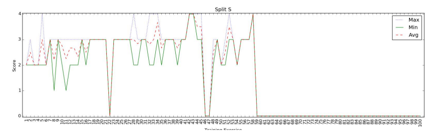


(b) Spin Recovery Statistics

Figure 120. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Spin Recovery

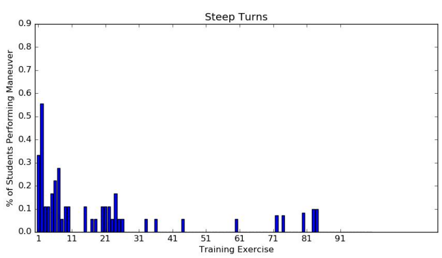


(a) Split S Occurrence

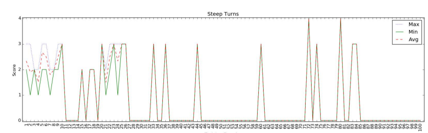


(b) Split S Statistics

Figure 121. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Split S

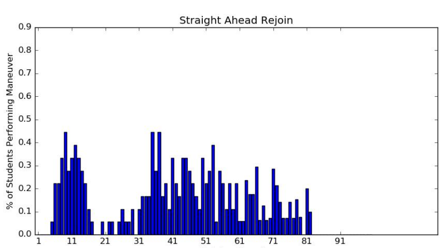


(a) Steep turns Occurrence

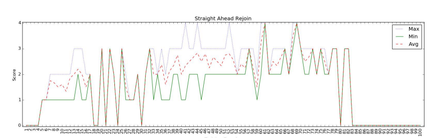


(b) Steep Turns Statistics

Figure 122. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Steep Turns

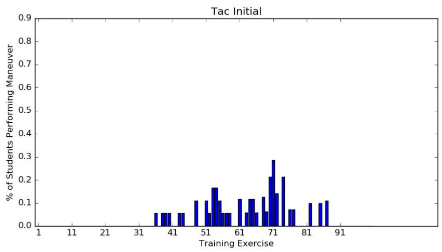


(a) Straight Ahead Rejoin Occurrence

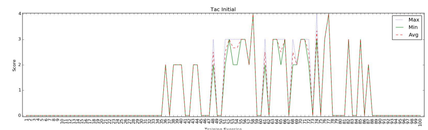


(b) Straight Ahead Rejoin Statistics

Figure 123. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Straight Ahead Rejoin

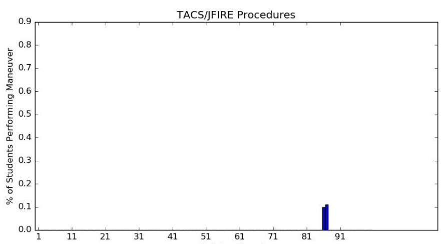


(a) Tac Initial Occurrence

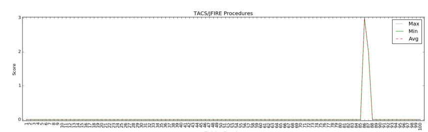


(b) Tac Initial Statistics

Figure 124. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Tac Initial

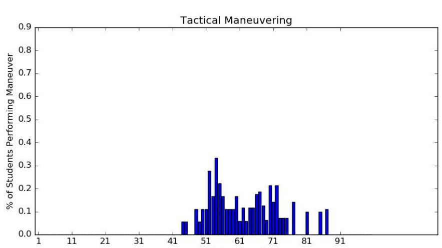


(a) TACS/JFIRE Procedures Occurrence

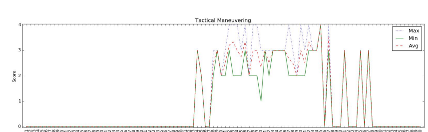


(b) TACS/JFIRE Procedures Statistics

Figure 125. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for TACS/JFIRE Procedures

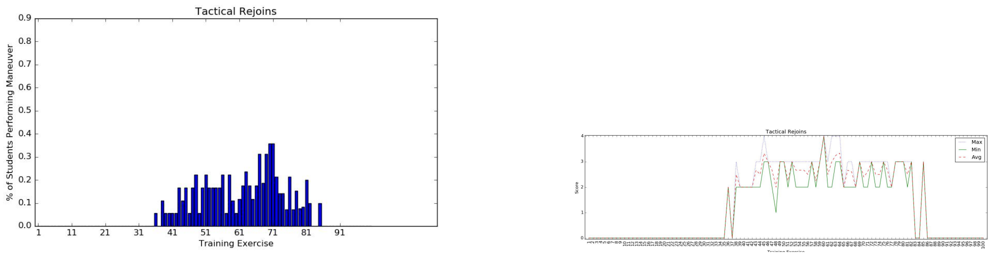


(a) Tactical Maneuvering Occurrence



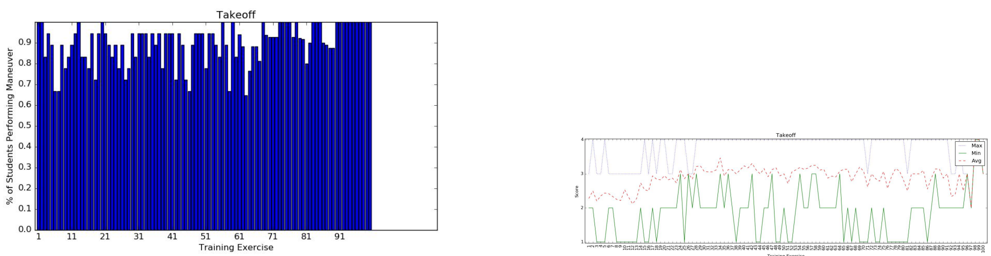
(b) Tactical Maneuvering Statistics

Figure 126. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Tactical Maneuvering



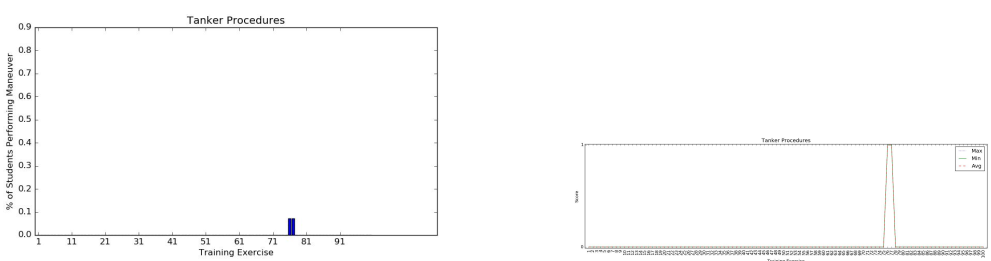
(a) Tactical Rejoins Occurrence (b) Tactical Rejoins Statistics

Figure 127. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Tactical Rejoins



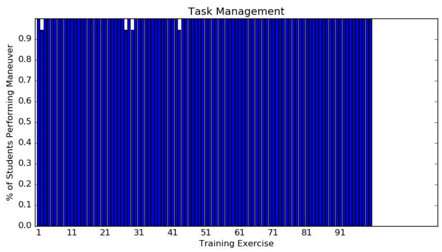
(a) Takeoff Occurrence (b) Takeoff Statistics

Figure 128. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Takeoff

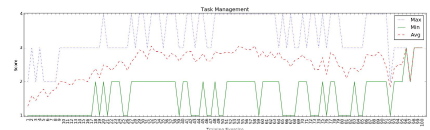


(a) Tanker Procedures Occurrence (b) Tanker Procedures Statistics

Figure 129. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Tanker Procedures

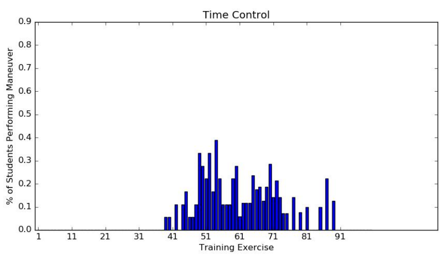


(a) Task Management Occurrence

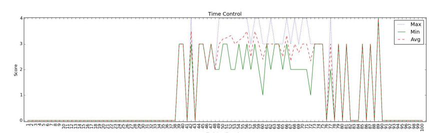


(b) Task Management Statistics

Figure 130. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Task Management

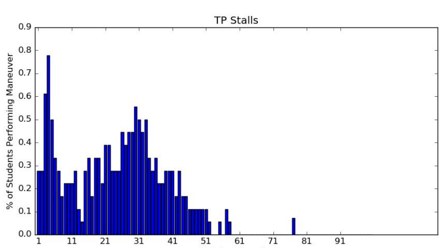


(a) Time Control Occurrence

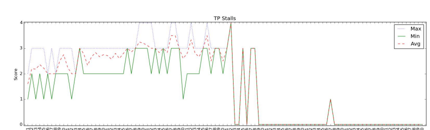


(b) Time Control Statistics

Figure 131. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Time Control

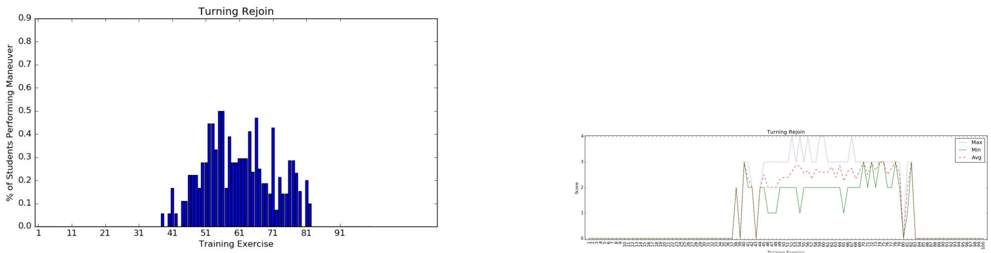


(a) TP Stalls Occurrence



(b) TP Stalls Statistics

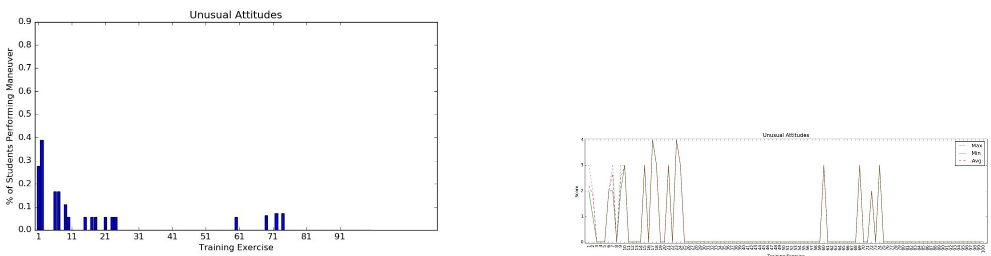
Figure 132. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for TP Stalls



(a) Turning Rejoin Occurrence

(b) Turning Rejoin Statistics

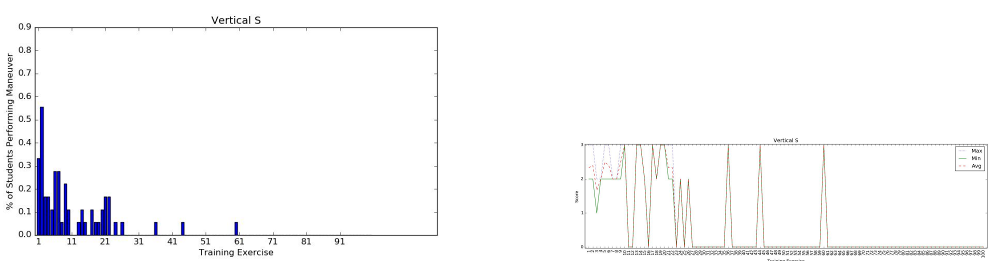
Figure 133. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Turning Rejoin



(a) Unusual Attitudes Occurrence

(b) Unusual Attitudes Statistics

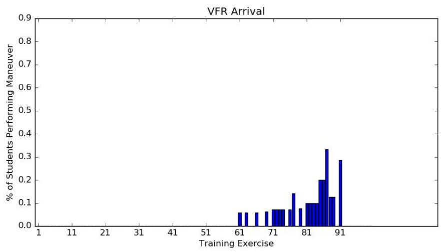
Figure 134. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for all events



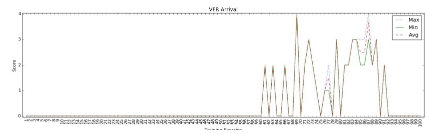
(a) Vertical S Occurrence

(b) Vertical S Statistics

Figure 135. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Vertical S

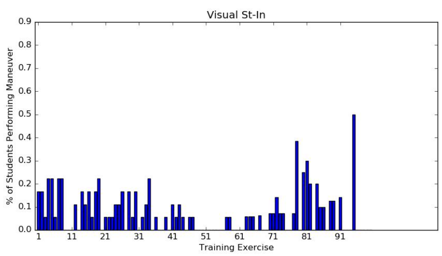


(a) VFR Arrival Occurrence

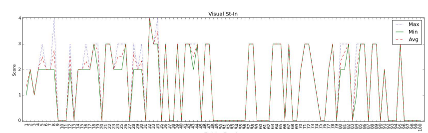


(b) VFR Arrival Statistics

Figure 136. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for VFR Arrival

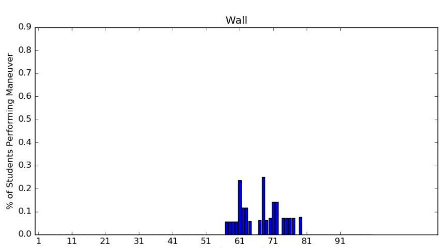


(a) Visual St-In Occurrence

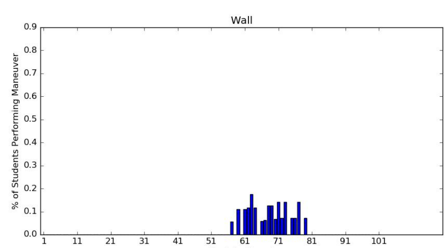


(b) Visual St-In Statistics

Figure 137. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Visual St-In

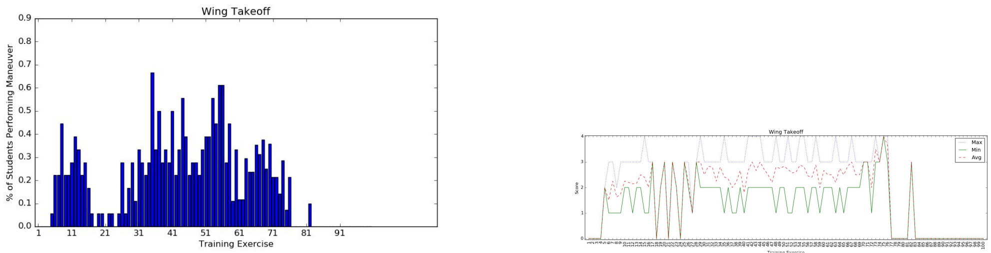


(a) Wall Occurrence



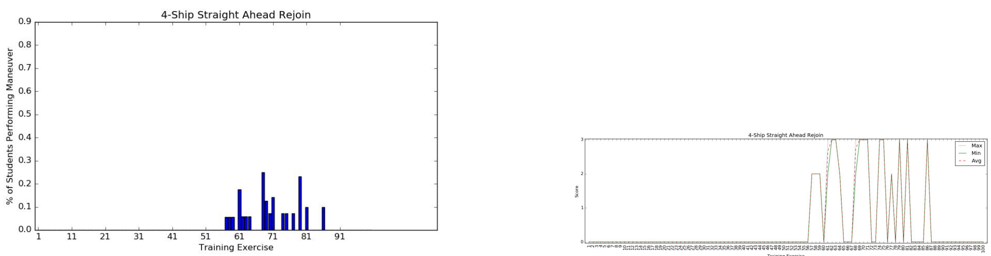
(b) Wall Statistics

Figure 138. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Wall



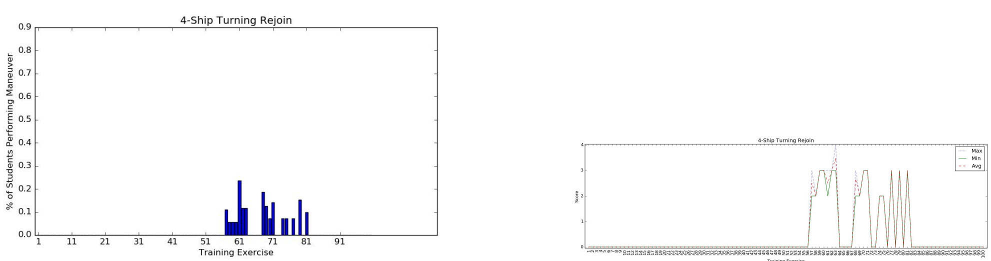
(a) Wing Takeoff Occurrence (b) Wing Takeoff Statistics

Figure 139. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Wing Takeoff



(a) 4-Ship Straight Ahead Rejoin Occurrence (b) 4-Ship Straight Ahead Rejoin Statistics

Figure 140. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for 4-Ship Straight Ahead Rejoin



(a) 4-Ship Turning Rejoin Occurrence (b) Turning Rejoin Statistics

Figure 141. Probability of event occurrence (a) alongside descriptive evaluation data (b) given training exercise for Turning Rejoin

Appendix B: Bullet Background Paper

BULLET BACKGROUND PAPER
ON
AUTOMATED FLIGHT PLAN RECOMMENDATION

PURPOSE

This paper provides key takeaways concerning novel research conducted to create the first ever automated training flight plan development process for the Pilot Training Next Program (PTN). The highlights include the network, process, and value provided by this research.

NETWORK

- The network includes all involved in findings implementation: AETC, AFWERX-Austin, Pilot Training Next, undergraduate pilot training Instructor Pilots (IPs), and future students

PROCESS

- Current process: IPs spend hours each day planning which flight events should be evaluated in a pilot training students next training exercise.
- Realized process: Applied industry leading Artificial Intelligence techniques to automatically propose a set of flight events for IP approval or alteration. First such system developed.

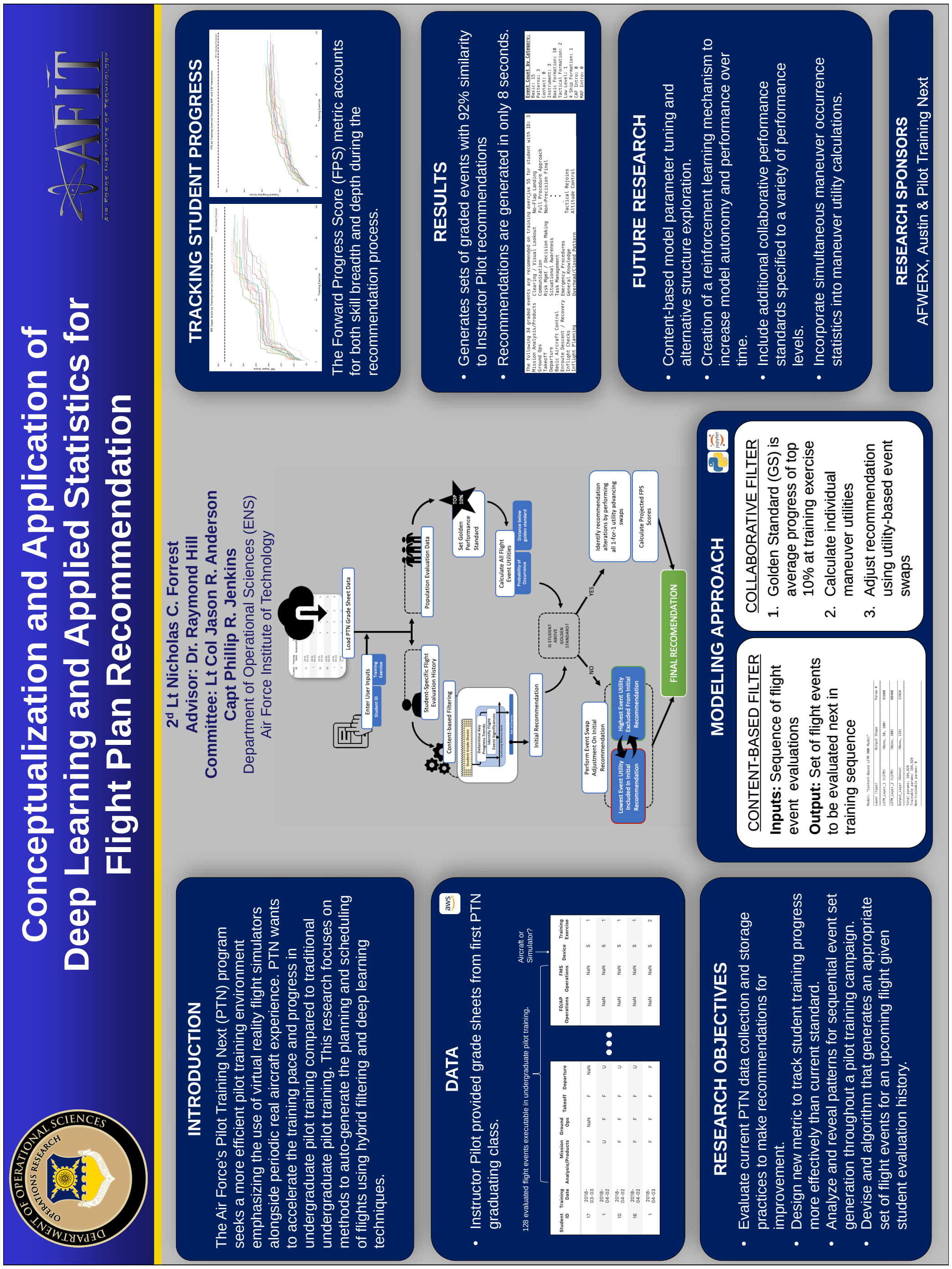
VALUE

- Flight planning time reduced to seconds, allowing IP time reallocation to personalized student development. Implementation projected to save 500+ IP hours per PTN class.
- Development of new training progress metric allows for unambiguous performance tracking and comparison between pilot training students.
- Guidance for establishing an effective data environment provides a foundation for future data implementation.
- Applications of automated flight planning go beyond PTN, can provide competent and quality training guidance in other Formal Training Units, non-AETC training environments such as USAFA or ROTC detachments, and operational flight optimization.

CONCLUSION

The research conducted on flight plan automation found that Artificial Intelligence techniques can be implemented to quickly and effectively provide instructor pilots with flight event recommendations for an upcoming flight. Insights gathered from current operations motivated the proposal of a more effective data environment and student performance measurement.

Appendix C: Project Poster



Bibliography

1. J. Mattis, “Summary of the 2019 National Defense Strategy of The United States of America,” tech. rep., United States of America Department of Defense, Washington, D.C., 2018.

2. M. Coffman, “Transcript of hearing before the subcommittee on military personnel of the committee on armed services, house of representatives, one hundred fifteenth congress, first session on miltary pilot shortage,” 2017.

3. Alderson Court Reporting, “Stenographic Transcript Before the Subcommittee on Readiness and Management Support COMMITTEE ON ARMED SERVICES UNITED STATES SENATE HEARING TO RECEIVE TESTIMONY ON UNITED STATES AIR FORCE READINESS,” 2018.

4. P. Doherty, “T-6A Primary Pilot Training,” Headquarters Air Education and Training Command, Department of the Air Force, 2019.

5. B. Oliver, *Optimizing the Undergraduate Pilot Training Scheduling Process.* PhD thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio, 2007.

6. M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, “Collaborative Filtering Recommender Systems,” *Foundations and Trends in Human-Computer Interaction*, vol. 4, no. 2, pp. 81–173, 2011.

7. Y. Koren, R. Bell, and C. Volinsky, “Matrix Factorization Techniques for Recommender Systems,” tech. rep., IEEE Computer Society, 2009.

8. B. N. Miller, “GroupLens: An Open Architecture for Collaborative Filtering,” tech. rep., Pennsylvania State College of Information, Science, and Technology, 1995.

9. Y. Koren, “Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model,” tech. rep., Las Vegas, Nevada,USA, 2008.

10. G. Linden, B. Smith, and J. York, “Amazon.com Recommendations,” *IEEE Internet Computing*, 2003.

11. Y. Z. Yao and Y. R. Xu, “Parameter analysis of particle swarm optimization algorithm,” *Harbin Gongcheng Daxue Xuebao/Journal of Harbin Engineering University*, vol. 28, no. 11, pp. 1242–1246, 2007.

12. R. Eberhart and James Kennedy, “A New Optimizer Using Particle Swarm Theory,” *Sixth International Symposium on Micro Machine and Human Science*, vol. 0-7803-267, pp. 39–43, 1999.

13. B. Sanghavi, R. Rathod, and D. Mistry, “Recommender Systems-Comparison of Content-based Filtering and Collaborative Filtering,” *International Journal of Current Engineering and Technology*, vol. 4, no. 5, 2014.

14. J. Pennington, R. Socher, and C. Manning, “Glove: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Stroudsburg, PA, USA), pp. 1532–1543, Association for Computational Linguistics, 2014.

15. J. Xie and Q. Wang, “Benchmark machine learning approaches with classical time series approaches on the blood glucose level prediction challenge,” *CEUR Workshop Proceedings*, vol. 2148, pp. 97–102, 2018.

16. F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, eds., *Recommender Systems Handbook*. New York, NY: Springer Science+Business Media, LLC, 2011.

17. A. Srivastava, P. K. Bala, and B. Kumar, “New perspectives on gray sheep behavior in E-commerce recommendations,” *Elsevier Journal of Retailing and Consumer Services*, vol. 53, pp. 1–11, 2019.

18. M. Dhanda and V. Verma, “Recommender System for Academic Literature with Incremental Dataset,” in *Twelfth International Multi-Conference on Information Processing-2016*, vol. 89, pp. 483–491, Elsevier, 2016.

19. T. Gao, J. Chen, V. Chenthamarakshan, and M. Witbrock, “A Sequential Set Generation Method for Predicting Set-Valued Outputs,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 2835–2842, 2019.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
26–03–2020	Master's Thesis	Sept 2018 — Mar 2020

4. TITLE AND SUBTITLE

Conceptualization and Application of Deep Learning and Applied Statistics for Flight Plan Recommendation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Nicholas C. Forrest

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management
Department of Operational Sciences (AFIT/ENS)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-20-M-147

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFWERX Austin Innovation Hub
701 Brazos St
Austin, TX 78701
Email: kyle.palko@afwerx.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

AFWERX

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Air Forces Pilot Training Next (PTN) program seeks a more efficient pilot training environment emphasizing the use of virtual reality flight simulators alongside periodic real aircraft experience. The PTN program wants to accelerate the training pace and progress in undergraduate pilot training compared to traditional undergraduate pilot training. Currently, instructor pilots spend excessive time planning and scheduling flights. This research focuses on methods to auto-generate the planning of in-flight events using hybrid filtering and deep learning techniques. The resulting approach captures temporal trends of user-specific and program-wide student performance to recommend a feasible set of graded flight events for evaluation in a students next training exercise to improve their progress toward fully qualified status.

15. SUBJECT TERMS

Pilot Training NEXT, Recommender Systems, Flight Planning, Deep Learning, Applied Statistics, Neural Networks, Recurrent Neural Networks

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
U	U	U	U	143	Dr. Raymond R. Hill, AFIT/ENS
					19b. TELEPHONE NUMBER *(include area code)*
					(937) 255-6565, x7469; rhill@afit.edu