An Examination of Potential Training Regression Recognition Algorithms for Pilot Training Next

THESIS

Alex Gaines, 1st Lt, USAF

AFIT-ENS-MS-21-M-160

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-21-M-160

AN EXAMINATION OF POTENTIAL TRAINING REGRESSION RECOGNITION ALGORITHMS FOR PILOT TRAINING NEXT

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

Alex Gaines, BS

1st Lt, USAF

25 March 2021

AFIT-ENS-MS-21-M-160

AN EXAMINATION OF POTENTIAL TRAINING REGRESSION RECOGNITION ALGORITHMS FOR PILOT TRAINING NEXT

THESIS

Alex Gaines, BS
1st Lt, USAF

Committee Membership:

Dr. R. R. Hill
Chairman

Dr. A. Reiman
Member

Dr. P. Jenkins,
Member

AFIT-ENS-MS-21-M-160

Abstract

The initiative to reduce the Air Force's serious pilot shortage led to the Pilot Training Next (PTN) program. Under PTN, student pilots progress at an individual rate while making increased use of simulator-based training resources. A previous thesis used data from the first PTN class to conceptualize and prototype a student training flight scheduler. This scheduler did not consider training events required to bring students back to achieved levels of performance if in fact that student performance had regressed. This thesis examines three classes of PTN student data to determine whether student regression in training progression can be detected. A visual, a statistical, and a machine learning-based method are examined and found to not predict training regression in PTN student pilots.

This is dedicated to my family and all others that have continued to support my academic endeavors.

Acknowledgements

I want to thank my fellow students and my thesis committee for their continual support throughout this graduate program.

Alex Gaines

Table of Contents

Page

List of Figures

List of Tables

AN EXAMINATION OF POTENTIAL TRAINING REGRESSION RECOGNITION ALGORITHMS FOR PILOT TRAINING NEXT

I. Introduction

The Air Force has been unable to train a sufficient number of pilots to meet operational demands Robbert et al. (2015). The Pilot Training Next (PTN) program seeks to address the problem through personalized, simulator-based training. Students in the PTN program have around the clock access to simulators and do not follow a rigid curriculum. This allows students to progress at their own pace and finish training sooner than they would in the traditional undergraduate pilot training (UPT) program.

However, this increased personalization facilitated by PTN requires instructor pilots (IP) to spend an inordinate amount of time on administrative tasks such as developing and organizing flight plans for students. It also induces subjectivity in evaluations because students may be assessed by different instructors or on different dates. An automated IP system, AutoGradebook, is under development to help rectify these issues.

Previous research conducted by Forrest (2020) focused on automatically recommending training programs through AutoGradebook in order to train pilots more efficiently. This work builds upon the previous research by describing approaches to identify students who may struggle with pilot training. If these students can be identified early in training then IPs may intervene with training regimens intended to rapidly develop their skills to a baseline level and prevent regressions in performance.

Training regression is defined here as occurring when pilot student proficiency in

a task recedes. An IP may notice such regression and plan subsequent training events to reverse the regression. An automated regression notification system can detect the regression in training and cause the automatic training program scheduler to include requisite training events to counter the regression. The primary research objective in this work is to determine whether such an automated regression notification system is possible for the PTN program. More specific research questions include examining which regression detection methods might be useful for PTN, determining whether the data collection methods and data employed in PTN support the regression detection task, and what are the paths to implementation of a regression detection tool.

The remainder of this thesis is organized as follows. Chapter II provides a literature review of topics pertinent to recommending and scheduling training programs and detecting regressions in performance. Various automated systems, statistical methods, and psychological phenomena are discussed. Chapter III describes the analytical techniques that were applied to the PTN data as well as techniques that could be employed with additional data. Chapter IV provides insights and suggestions to improve the PTN program.

II. Literature Review

2.1 Individualized Training

Individualized training programs are tailored to the needs of each student. Students in individualized training programs spend less time on topics they are already proficient in. This time savings can provide them with more time to develop their weaker skills or allow them to proceed to more advanced topics. These individualized programs can allow strong students to graduate earlier than they would in traditional training programs while also increasing the likelihood that weaker students are able to complete training.

Individualized education systems have been used to help students succeed in American schools for decades Betrus (1995). They have also been embraced by the private sector to help meet a variety of ends such as personal fitness training. Technological advancement has made individualized education and training even more viable by reducing the amount of instructor engagement that is required or even eliminating it entirely in some cases. Various commercial language learning applications, such as Duolingo use automated methods to suggest and create individual study plans Settles and Meeder (2016). Individualized training programs are also becoming more common throughout the government and military with the goals of reducing costs and improving training quality Manacapilli et al. (2011).

The main drawbacks of individualized training are the amount of time required by instructors and the number of instructors required for multiple students Betrus (1995). An instructor of a traditional education program may serve several students while an instructor for an individualized program may only have bandwidth for a few students. This means that an individualized training programs may have lower throughput than a traditional program, despite their potential to train students more

quickly. Automated systems can mitigate this problem by assisting instructors in the development of training programs or identifying student performance patterns.

2.2 Cognitive Psychology

Various psychological phenomena may be pertinent to developing individualized training regimens. The learning curve, the forgetting curve, and the spacing effect were first described by Ebbinghaus in 1885 Ebbinghaus (2013). Learning curves relate an individual's proficiency at a task to the amount of time spent on the task. Forgetting curves relate an individual's memory of a task to the time elapsed when they are not training on the task. Memory retention declines exponentially according to the forgetting curve model. However, these forgetting curves become less steep with increasingly spaced repetitions of learning events. This phenomenon is the spacing effect.

Forgetting curves and the spacing effect form the basis of many learning programs and memory models. The Leitner system is commonly used method for studying flash cards wherein material that is commonly missed is reviewed more frequently than material that is well known Leitner (2008). Duolingo, a language learning application and company, has developed an individualized model to estimate the half life of words or concepts for their students Settles and Meeder (2016). Others have developed models to produce review schedules that maximize an individual's learning rate Reddy et al. (2016). Air Force Research Laboratory (AFRL) researchers have developed a model to prescribe training regimens to meet a variety of goals, such as achieving a target level of performance by a specified date or minimizing the number of training events required to maintain proficiency through a specified date Jastrzembski et al. (2013). Adaptations of these models may be particularly useful for tracking student proficiency and developing pilot training regimens that deter regression in student

capabilities.

2.3 Machine Learning

Machine learning (ML) is a branch of artificial intelligence concerned with designing systems that are able to learn from experience Kirk (2017). In contrast to mathematical and statistical models which are built on fixed equations, machine learning models may not assume a particular relationship between inputs and outputs. This allows use of machine learning algorithms for a variety of real-world tasks that are too complex for purely statistical methods. There are three types of machine learning algorithms: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning is used to make inferences about an unobservable item of interest (output) based on a set of related observations (inputs). The goal of a supervised learning problem is to create a function that accurately maps inputs to the output. There are two types of supervised learning problems which are based on the data type of the output. Classification problems deal with categorical output while regression problems (not to be confused with regression in performance) deal with continuous output.

Unsupervised learning is used to identify the hidden patterns and salient features within a data set, yielding a greater understanding the data. Common unsupervised learning goals are clustering and dimensionality reduction. Clustering is used to partition data into related groups (clusters). Items within a cluster have a stronger relation to each other than to items in any other cluster. Dimensionality reduction techniques are used to reduce the size of a data set in a way that retains information that will be conducive for future analyses. Dimensionality reduction may also be used for data visualization, making large data sets easier to interpret.

Reinforcement learning is used to teach autonomous agents to perform well in

sequential decision-making problems under uncertainty. Performance at each decision epoch is quantified via a reward, a function of the agent's decision and its current environment. The overall goal is for the agent to learn a decision-making policy that maximizes the total reward it accrues over all decision epochs. In other words, the agent must learn the long-term impacts of its decisions to be successful.

Of the ML applications discussed, both types of supervised learning algorithms appear pertinent to developing systems to mitigate performance regressions during PTN. Using supervised learning to classify students into learning profiles early in training should help IPs develop better training programs. If students can be classified into categories such as "likely to wash out", "likely to regress", "late bloomer", or "high performer" then IPs may intervene where appropriate to ensure that students maintain standards of performance. A more direct approach is using supervised learning to model students' performance scores over time as a regression problem. If accurate performance models can be developed, then these models may be used to predict when a student starts to regress and the particular subjects or skills that the students are likely to regress in.

2.4 Recommender Systems

The space of training options for a student in PTN is large and difficult to evaluate. There are over 100 individual maneuvers a student pilot may be graded on a smaller subset of these maneuvers comprise each individual training exercise. Instructor pilots manually select these smaller sets of maneuvers based on each student's performance. This is an inordinately time-consuming process, so an automated system to suggest training events would be highly beneficial. Various computer-based options such as recommender systems can integrate previously recorded user data to generate suggestions automatically Ekstrand et al. (2011). Such recommender systems have

been utilized by companies like Amazon to immense commercial success Koren et al. (2009) Smith and Linden (2017). The incorporation of recommender system into the AutoGradebook should substantially reduce the amount of time instructors spend on creating and evaluating training profiles, giving them more time to train and mentor students.

A recommender system, also sometimes called a recommendation engine, is an information system that examines data associated with some performance of interest and uses that data to make recommendations for future performance. For instance, when using Amazon to purchase books, an underlying recommendation engine will suggest other, similar books to consider purchasing. As computer systems have improved and pattern recognition algorithms have matured, recommender systems have grown in accuracy and use.

A common recommendation generation method is collaborative filtering which utilizes prior user behavior to predict their future actions. It is based on the simple assumption that similar users should be given similar recommendations. Two successful collaborative filtering approaches are latent factor models and neighborhood models. Latent factor models are able to capture trends indicated weakly by many data points while neighborhood models are able to identify trends indicated strongly by only a few data points. A combination of these approaches may also be used to improve predictive accuracy Koren (2008).

Other data driven techniques have been shown to be practical in industry. Least squares and k-nearest-neighbor techniques are two commonly applied, simple, yet powerful methods for data classification and prediction and form the basis for other more advanced models Hastie et al. (2009). More advanced techniques such as machine learning may also be useful. Neural networks in particular are effective for modeling and predicting time-series data generated by biological entities Patterson

and Gibson (2017).

A combination of classification, recommendation, prediction approaches seems appropriate for the AutoGradebook due to its unique operational environment. Typical commercial recommender systems seek to identify and remove outlier users from the data as their presence can lower the overall quality of recommendations Srivastava et al. (2020). Rather than simply removing students with atypical performance patterns from the data pool, the AutoGradebook should classify them separately and generate individualized recommendations and predictions. Incorporating these capabilities could enable instructors to detect students' skill deficiencies sooner and facilitate the development of a more effective individualized pilot training program.

III. Methodology and Analyses Employed

3.1 Data Description

Each row of the raw PTN data consists of columns which indicate the student pilot (SP), the flight maneuver being conducted, the category of the maneuver, whether the flight was conducted live or in a simulator, the date of the flight, and the SPs grade for the maneuver as assessed by the IP. An abbreviated example of the raw data format is shown in Table 1.

Table 1. PTN Raw Data Example

Student	Maneuver	Flight Type	Date	Grade
INDIE11	Aileron Roll	Flight	2019-04-10	S
INDIE11	Barrel Roll	Flight	2019-04-10	G
INDIE11	Immelmann	Sim	2019-04-12	U
INDIE27	Aileron Roll	Sim	2019-04-19	E
INDIE27	Barrel Roll	Sim	2019-04-25	F
INDIE27	Immelmann	Flight	2019-04-25	G

The maneuvers and their categories are shown in Table 2. As easily discernible from Table 2, pilots train for many maneuvers and clearly not every maneuver is found in every training event.

Table 2. PTN 2 Maneuvers by Category

Category	Maneuvers

4 Ship Formation (4S)	Box Formation, 4S Tactical Maneuvering, 4S Mutual Support / Flight Integrity, 4S BD Check, Wall, 4S Fingertip, 4S Admin, 4S Turning Rejoin, Offset Box, 4S Straight Ahead Rejoin, Fluid 4
Admin	Task Management, Fuel Awareness / Management, Inflight Checks / Checklist Usage, Mission Planning / Briefing / Debriefing, Risk Mgmt / Decision Making, General Knowledge, Takeoff, Inflight Planning, Clearing / Visual Lookout, Basic Aircraft Control, Flight Discipline, Situational Awareness, Ground Ops, Emergency Procedures, Divert/Contingency Planning, Cross-Check, Communication, Mission Analysis/Products, Mission Management, Enroute Descent / Recovery, Departure

Basic Formation (2S)	2S G-Warmup / Awareness, 2S Breakout (Wing), Formation Landing (Both), 2S Echelon (Wing), 2S Fighting Wing (Wing), Route (Wing), Flight Split (Wing), Extended Trail (Wing), Lead Platform, Formation Approach (Both), Instrument Trail, Extended Trail (Lead), 2S Position Change, Lost Wingman (Both), Straight Ahead Rejoin, Overshoot, Pitchout (Both), 2S Interval Takeoff, Turning Rejoin, 2S Battle Damage Check, Wing Takeoff, 2S Fingertip (Wing), Flt Integrity / Wingman Consideration
CAF Advanced Air to Air Operations	Intercept to Position of Advantage, BVR/WVR Mutual Support, Engaged / Defensive / Mutual Support Maneuvering, Roles Establishment and Contract Adherence
CAF Air to Surface Weapons Employment	SA Safe-Escape Maneuver, SA Pattern Ops: Diving Deliveries (Conventional), Tactical Weapons Deliveries - Unguided, Tactical Weapons Deliveries - Guided, SA Pattern Ops: Diving Deliveries (Tactical), SA Conventional Range, SA Tactical Range Proc, CAF Air to Surface Error Analysis

CAF Intro	CAF Fuel Awareness / Management, CAF Sim Single Engine Go Around, Weapons Employment, ROE Adherence, HUD OFF-Normal Landing, Sensor Set up (AA & AG Radar, TPod, DASS), CAF Heavy Weight Touch and Go, FENCE, Programmable Armament Control System Set up, CAF Sim Single Engine St-in Landing, CAF Heavy Weight Considerations
CAF Introduction to Air to Air Operations	Heat to Guns Maneuvering Offensive, Long Range Offensive, CZ Recognition, Reversal Scissors Ex, Heat to Guns Maneuvering Defensive, HA BFM Flt Analysis, Air to Air Weapons Employ, Jink Exercise, HA Butterfly Setups, Short Range Offensive, Long Range Defensive, 1/4 Plane Ex, AA WEZ Recognition, Maneuver Mechanics/Execution, Range & Pursuit Curve Ex, Turn Circle Analysis, Short Range Defensive, Advanced Handling, HA Lead Turn Exercise, Perch Setups
CAF Surface Attack Tactical Operations	Surface Attack Prioritization, Air to Ground 2-ship Mutual Support, Low Alt Formation Mutual Support / Contract Adh, CAF SA Threat Reactions, Low Alt Tactical Holding, CAF TACS/JFIRE Procedures, First Run Attack, Surface Threat Awareness, Low Alt Systems Reprogramming

Contact/AHC	Contact Recoveries, Immelmann, TP Stalls, Barrel Roll, Slow Flight, Pitchback / Sliceback, Cuban Eight, ELP Stalls, Power On Stalls, Stall Awareness / Recoveries, G-Awareness, Lazy Eight, Cloverleaf, Loop, Spin Recovery, Aileron Roll, Split S
Instruments	Missed Approach, Unusual Attitudes, ASR/PAR Approach, Night Navigation, Precision Final, Holding, Non-Precision Final, Vertical S, Non-Precision Final (VOR/TCN), Intercept / Maintain Arc, Penetration Approach, RNAV Final, Steep Turns, Night Landing, Off Station Approach, Intercept / Maintain Course, Fix to Fix, Circling Approach, Full Procedure Approach
Low Level (LL)	LL Altitude Control, LL Lead Change, Course Entry, Low Alt Maneuvering, Route Abort / Low Level Exit, Checkpoint ID, LL Course Mx, LL Fighting Wing / Wedge, Ridge Crossing, LL GPS Integration, LL Tactical Maneuvering, Time Control
MAF/SOF Intro	Crew Coordination, MAF/SOF VFR Arrival, Pilot Monitoring, Tanker Procedures, Airdrop Procedures, Crew Resource Management, MAF/SOF Mission Management

Patterns	Go-Around, Pattern Night Landing, Overhead/-Closed Pattern, Landing, Visual St-In, No Flap Landing, SFL, Emergency Landing Pattern
Tactical Formation (TAC)	FM Level 1, Delayed 90, Shackle, 2S Tac Lead Platform, Tac Position (Wing), 2S Tac Straight Ahead Rejoin, 2S Tac Position Change, Cross Turn, 2S Tac Shackle, Fluid Maneuvering, 2S Tac Turning Rejoins, Fluid Turn, Delayed 45, 2S Tactical Maneuvering, 2S Tac Hook Turn, Tac Initial, FM Level 2
VFR Flight	VFR Arrival, VFR Navigation

Students are graded on a four-letter system by instructor pilots with a minimum score of Good being required to pass any maneuver. These grades and their numerical equivalents are shown in Table 3. The numerical equivalents are used to compute a metric called the cumulative maneuver item file (MIF) by summing the maximum score a student has achieved for each maneuver. This metric is used by IPs to assess the overall progress of students during pilot training. Using the Cumulative MIF as the primary measure of progress introduces problems with respect to training regression detection and mitigation goals. These are discussed in the next chapter.

Table 3. PTN Grading Scheme

Grade	Meaning	Points
E	Excellent	4
G	Good	3
F	Fair	2
U	Unsatisfactory	1
NG	Not Graded	0

The data set was partitioned by student and transformed with a pivot table.

This yielded student data tables with a row for every day the student completed a training exercise and columns for their scores in each maneuver conducted in that training exercise. Thus, each maneuver column is a time series of scores. However, the measurement intervals for these series are irregular due to students advancing at their own pace and the fact that students complete only a subset of the maneuvers during any training event. The implications of this are discussed below. An example of the student data table format is shown in Table 4. Note that students do not perform every maneuver during particular training events. There may also be gaps of varying length between maneuvers occurring within a students' training events.

Table 4. PTN Transformed Data Example

Date	Aileron Roll	Barrel Roll	Immelmann	Landing	Takeoff
2019-05-01	0	3	2	3	3
2019-05-04	4	3	0	2	4
2019-05-06	4	4	3	4	3
2019-05-09	3	0	3	3	3
2019-05-10	3	0	0	3	3
2019-05-15	0	3	4	3	2

PTN does appear to have structure with respect to the scheduling of graded training events, despite the self-paced nature of student progression through pilot training. Students typically spend the first two months of training on maneuvers from the Admin, Contact/AHC, Instruments, and Patterns categories. Admin maneuvers are conducted during nearly every training event for the duration of training campaign. Maneuvers from the latter three categories mentioned are basic skills that are built upon by maneuvers from the remaining categories. The more advanced maneuver categories are introduced after students demonstrate competence in the basic sets

of maneuvers. Some students were not introduced to certain advanced maneuver categories which suggests that students are classified into different training groups (perhaps for specialized pilot training) midway through PTN.

3.2 Early Insights

This data set has various issues that inhibit analysis and model building for the current purposes associated with detection of training regression. These issues are due to the metrics used by PTN leaders and PTN's unique operating environment when compared to traditional training programs. These issues are discussed here and recommendations to address them are discussed in Chapter 4.

First, maneuvers are scored on an ordinal scale rather than an interval or ratio scale. This means differences between scores are not quantifiable. For example, it is not clear whether moving from Excellent to Good is as significant as moving from Fair to Unsatisfactory despite these both being one point decrements. Additionally, the low granularity and subjectivity of this scoring system makes it difficult to determine whether any one point change in score can be attributed to a meaningful change in the student's performance level. These factors make determining whether performance regression has occurred difficult.

Next, the cumulative MIF metric does not necessarily correspond to a student's current level of performance since it is based on the maximum of all recorded scores for each maneuver. For example, if a student received an Excellent in Decision Making during week 10 and an Fair in Decision Making during week 20 then their cumulative MIF during week 20 would be based on the Excellent they received previously. Thus, cumulative MIF is not a useful indicator for regression recognition. MIF can also be ambiguous when used to compare students since MIF does not account for the breadth of training received. For example, two Excellents in separate maneuvers is

worth the same number of points as four Fairs. Thus, comparisons between students who have been exposed to differing numbers of maneuvers can be misleading.

Finally, the time series for each maneuver has missing data due to the nature of PTN and the large number of potential maneuver choices used to create each individualized training event. There are more than 150 individual maneuvers, so it is not possible to perform every maneuver during each training event. This yields a sparse time series which must be imputed in order to use certain multivariate time series analysis techniques. This data sparcity issue can be alleviated somewhat by grouping maneuvers by categories and/or excluding some maneuvers from analysis. However, series other than Admin, Patterns, and cumulative MIF are still sparsely populated.

The analytical techniques that were applied to the PTN data are discussed in the following section. These techniques were unable to provide meaningful predictions about a student's performance. Thus, they do not identify early indicators that a student may regress in performance. The discussion focuses on the shortcomings of the techniques that were applied and on ways that the PTN data compilation may be improved for future cohorts to facilitate the automatic detection of training regression. Modeling techniques that could be applied with supplementary data are discussed in Chapter 4.

3.3 Applying a Visual Analytical Technique

A simple approach using only cumulative MIF to stratify students is easy to implement and may help IPs identify potential wash outs. The plot in Figure 1 displays the cumulative MIF accrued by each student in PTN cohort 2 against the day of training. The median cumulative MIF for the students still in training on a given day was subtracted off to provide a better picture of their current standing in

the cohort. Students' scores are colored with a green to yellow to red gradient based on their final cumulative MIF accrued. Green is used to indicate the stronger scores, red the weaker scores, and yellow the scores more toward the average performer.

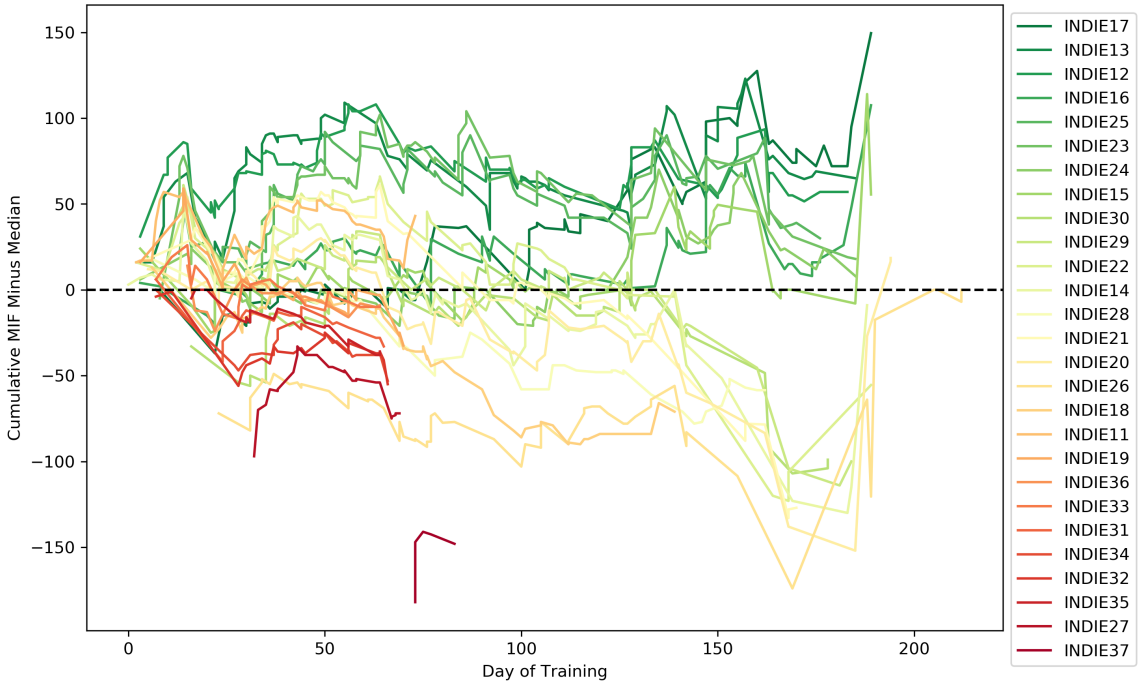

Figure 1. Adjusted cumulative MIF for each student in Cohort 2.

Stratification of the cohort is apparent within the first 50 days of training. A student that leaves training, or washes out, is indicated with the truncated line before PTN completion. The students who washed out were generally below the median during the first 50 days. Similarly, many of the students who finished PTN with the highest cumulative MIFs were above the median. However, there are a few notable exceptions to these trends. In particular, the student with the highest cumulative MIF at the end of training, INDIE17, spent most of the first 50 days of training below the median. Nevertheless, consistent performance below the median cumulative MIF at the beginning of PTN appears to be an early indicator that student is more likely to wash out.

While intuitive, this visual approach does not lend itself to automatic detection or prediction of training regression. It is not possible to determine whether a student is regressing from their total MIF alone because the metric is based on only their maximum recorded grades. As shown in Figure 2, a student's cumulative MIF over time is a monotonic increasing function characterized by periods of relative stagnation followed by sharp increases corresponding to the introduction of new maneuvers. A long period of decreases in the previous plot or stagnation in the following plot may be due to a student taking longer to master new concepts, rather than regressing in their capabilities on previously introduced concepts.

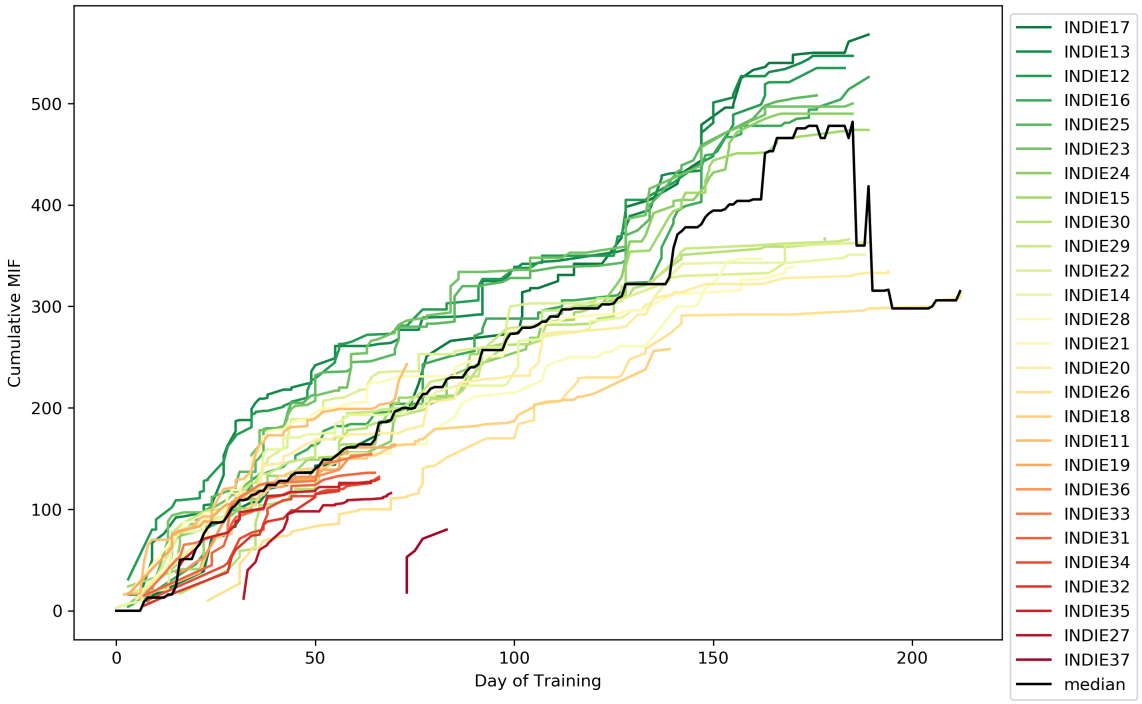

Figure 2. Student cumulative MIF over time for Cohort 2.

3.4 Applying Vector Autoregression

Detecting and predicting training performance regression is a complex task. If different maneuvers are interrelated, then a student's performance in one maneuver

may be indicative of their performance in others. Vector Autoregression (VAR) was chosen as a potential modeling technique because it relates multiple time series to each other and directly predicts variables values. More precisely, given a multivariate time series, the future values of each variable are modeled as linear combinations of previous variable values. Let y_t be a column vector of variable values at time t. A pth order VAR model is given by the equation $y_t = c + A_1 y_{t-1} + \ + A_p y_{t-p} + e_t$ where p is the number of previous time steps to include, A_i is a $k \times k$ coefficient matrix, c is a vector of constants, and e_t is a vector of error terms. A variable trending downward in successive predicted time steps could be an indicator that a student will regress in their training performance.

A VAR model was constructed for each student using the Statsmodels Python package. First, a student's data is partitioned into a training set and a validation set. Columns corresponding to maneuvers that were not performed during the training set are removed from both the training and validation sets since values for variables that were not observed cannot be modeled with VAR. Any missing data were imputed via linear interpolation. For example, suppose a student conducted Emergency Procedures during their first and third training events, but did not conduct them during their second training event. If the student's numerical scores for their Emergency Procedures were 3 and 4, respectively, then their score for their second training event would be imputed as 3.5. Next, a backward moving average based on five point window was used to smooth the data. Time step-wise differencing was used to remove the increasing trends in the data. This ensured that the variables are stationary, a requirement for VAR. A VAR model was then fit to the differences.

Forecasts from this VAR model are predictions on whether the students score for a particular maneuver will increase or decrease. Therefore, a large negative value or sequence of negative values in the forecast is a prediction of training performance

regression. These forecasts may be transformed back into scores by adding the final vector of scores in the training set to the cumulative sums of the forecast differences. The images in Figures 3 - 6 show the performance of a VAR model built on 60 training sessions of data.

Figure 3 shows the predictions for the differenced data against the actual differences. The VAR model did not yield meaningful predictions. The shape of the predictions is quite dissimilar to the observations. In particular, the model predicted that the student's score would decrease more frequently than it actually did.

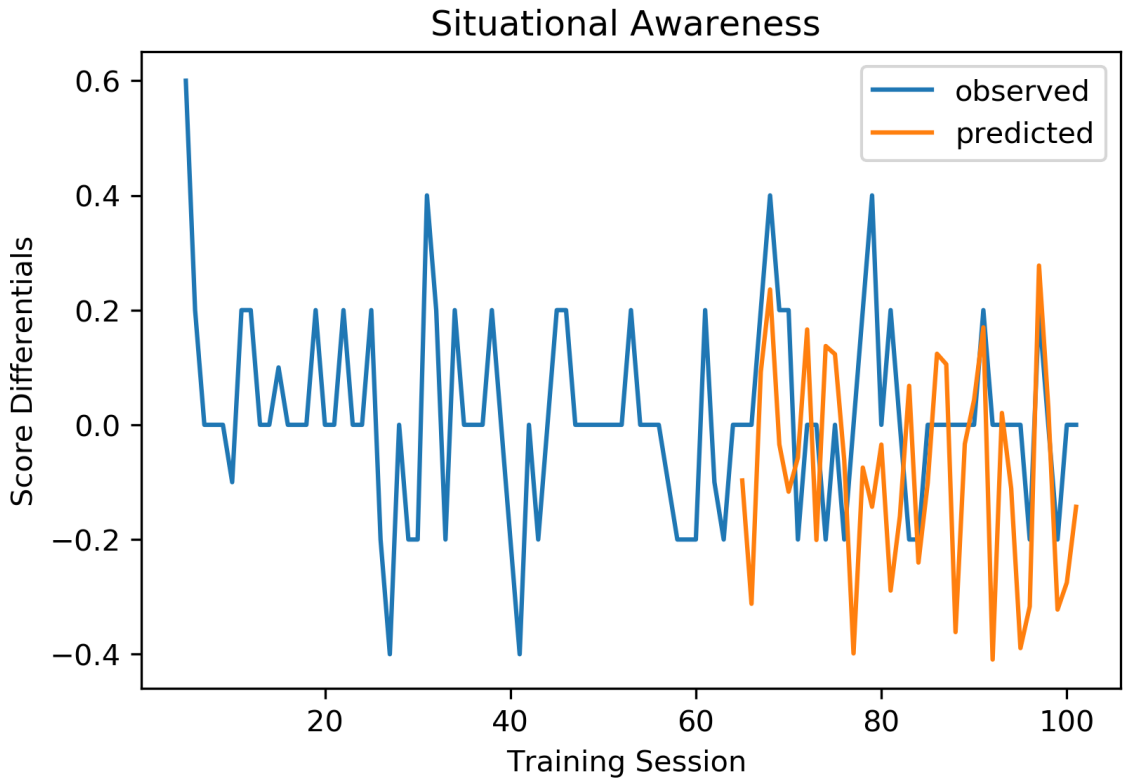

Figure 3. The VAR model was a poor fit for this student's Situational Awareness.

Figure 4 shows the predictions after converting back to scores. The errors in the predictions cascaded, producing an unrealistic series of scores. This was the typical behavior across all maneuvers; the converted predictions tended to have a large upward or downward trend.

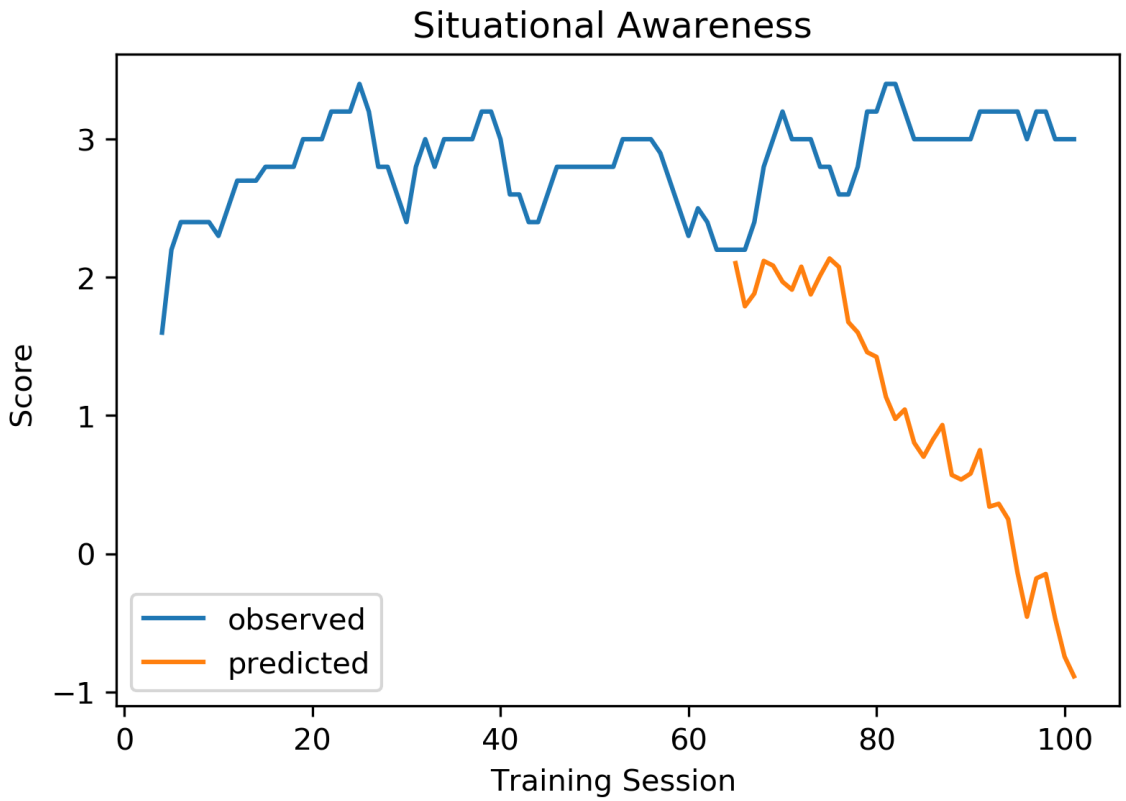

Figure 4. Converting the differences back into scores can produce infeasible results.

Figure 5 shows the predictions for Task Management. The predictions for this maneuver are much closer to the observations than those in Figure 3. Several of the early predictions are correct with respect to predicting an increase or decrease. However, the magnitudes of the predictions are still off in many cases.

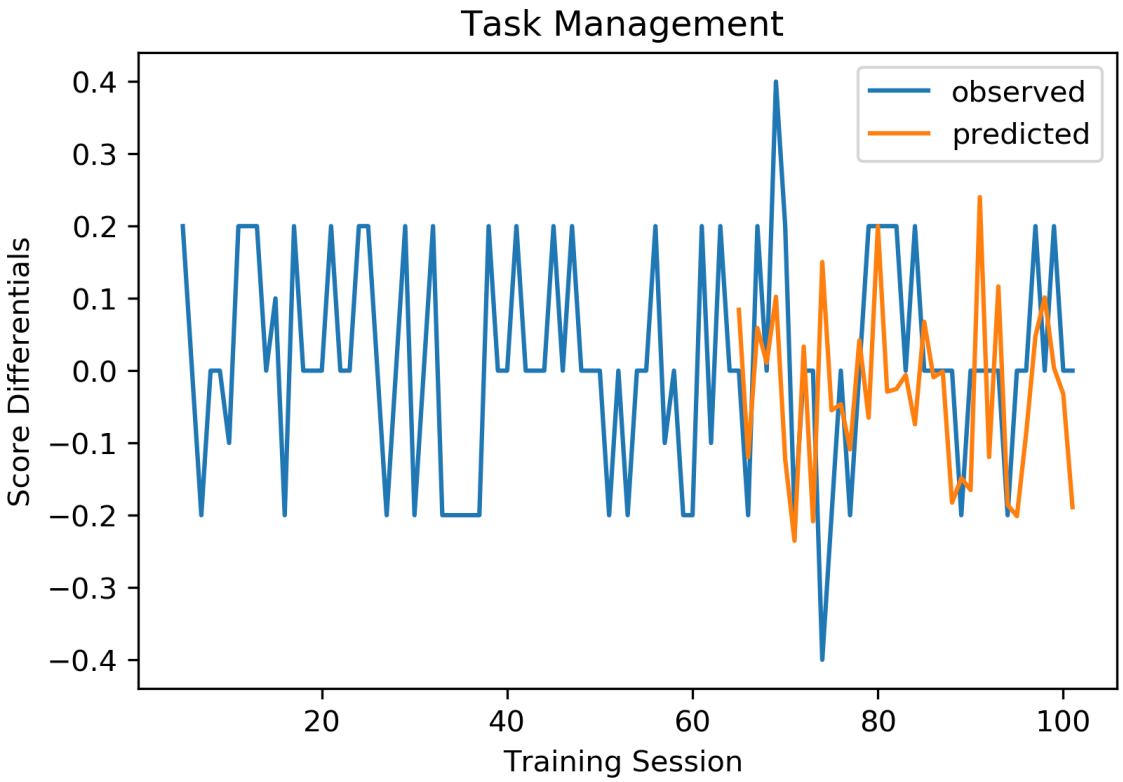

Figure 5. The VAR model predicted early score increases and decreases.

Figure 6 shows the converted Task Management predictions. The model predicted the decreasing scores that occurred between training sessions 65 - 75, and the inflection that occurred near training session 78. This is the only instance of a trend in this student's data being correctly predicted by the VAR model. However, this interpretation is dubious given the poor performance for every other maneuver.

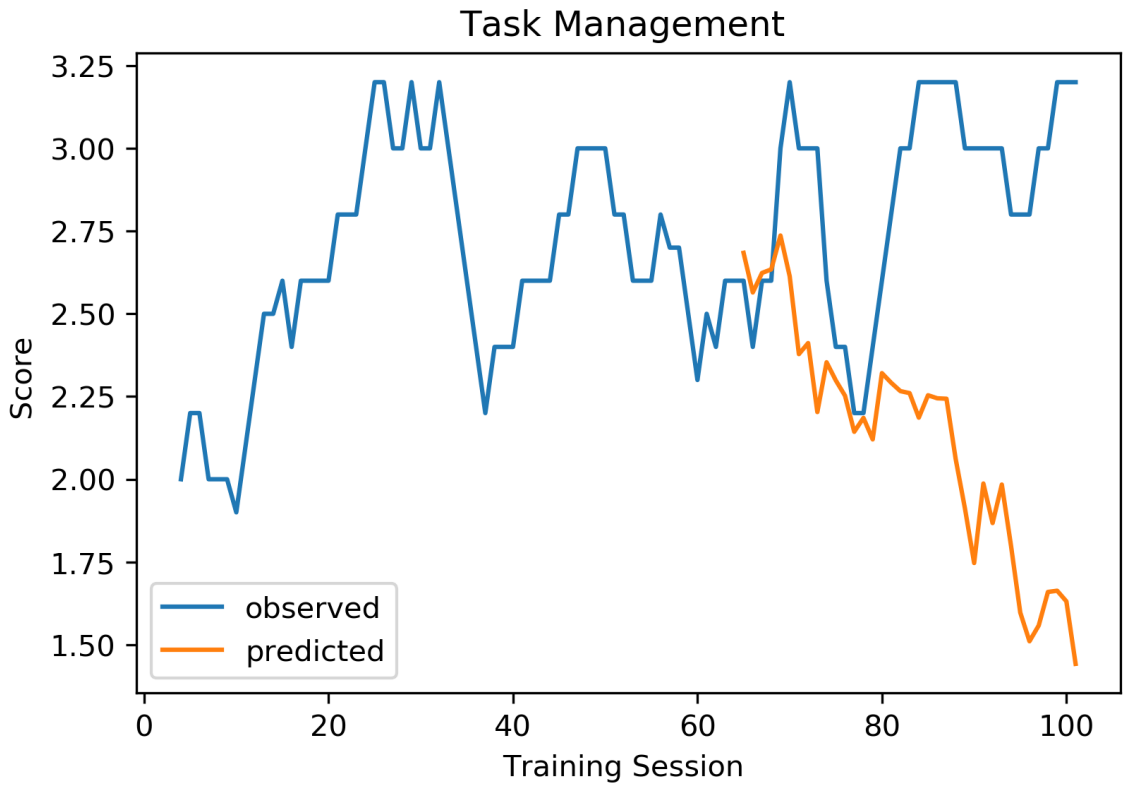

Figure 6. The model has dubious performance for this maneuver.

VAR had poor predictive performance for the PTN data set. There are several weaknesses and limitations of the VAR model which make it unsuitable for modeling human performance in the current context. Many students performed poorly on certain maneuvers during the first few months PTN, corresponding to the training portion of their data set, but improved dramatically by the end. This is reflective of the learning process. Students improve with practice and coaching. However, there is nothing to drive this behavior in the VAR model as it does not take studying, learning, or other human factors into account. Additionally, the sparsity and low granularity of the PTN data set makes identifying students' gradual knowledge and skill acquisition challenging to model and predict.

Another weakness of the VAR model is that the forecasts it produces are based solely on previous variable values. When forecasting several steps into the future,

later predictions are computed based on previous computed predictions. This causes errors to cascade and can lead to wildly inaccurate, or even implausible predictions over multiple time steps. However, the VAR model does not tend to predict large increases or decreases for a variable within a single time step. The combination of these factors makes accurately predicting training regression in the PTN difficult as maneuver scores can in fact change drastically between training events.

There are also practical concerns when using VAR and similar statistical regression models. A model must be constructed for every individual student pilot to generate performance forecasts. Then, additional processing on the forecasts must be used to determine whether performance regression is likely. This process is computationally expensive, although it may be worthwhile if a method to generate more accurate performance forecasts is found.

3.5 Applying a Long Short-Term Memory Recurrent Neural Network

The final analytical technique applied was a long short-term memory (LSTM) recurrent neural network (RNN). In contrast to the autoregressive models, RNNs can be used to predict values of various types of explanatory variables (e.g. categorical responses). RNNs also allow use of additional explanatory variables to predict the responses of interest. These features were leveraged to incorporate the gaps in manueuver consideration between successive training events and to make use of binary (pass or fail) predictions for maneuvers included in any of the training events.

Neural networks (NN) are machine learning algorithms that are loosely based on the hypothesized functioning of the human brain. They may be represented as a directed graph comprised of nodes which are organized in layers and weighted edges to connect the nodes. Every NN consists of a input layer, at least one hidden layer, and an output layer. Each node computes a weighted sum of its inputs then passes

the sum through a user-specified activation function to produce an output. These output values are then passed on as inputs for the next layer. The value(s) in the output layer is(are) a response prediction based on the input and the edge weights which are tuned algorithmically for a user specified number of epochs in a process known as training or learning.

RNNs contain edges which connect nodes back to nodes in the previous layer, allowing information persistence. This makes RNNs useful for sequential and time-series data. However, standard RNNs gradually forget what they have learned over several time steps, making them unsuitable for certain tasks. This problem is rectified by LSTM networks. LSTM networks are RNNs with additional components which help the network learn context based, long-term time dependencies from the input data.

LSTM models were constructed according to the following process based on using the Keras Python APIs. The first few data preparation steps are the same as those performed for VAR. First, student data are partitioned into a training and validation set. The partition is based on event timing; early data are in the training set and later data are for validation. As before, columns for those maneuvers not conducted during the particular training event are dropped. Missing data points were then imputed with linear interpolation. Two additional features were created to help make use of RNN capabilities and capture behavior that the VAR model could not capture. The first feature, "Training Lag", is the number of days since the previous training exercise. This feature was incorporated to implicitly model forgetting curves. The second additional feature dubbed "Skill" was intended to help model the increasing "Skill" at each time step and was computed as their cumulative MIF at each time step, normalized by dividing by 200. This normalization process produced a monotonic function with a similar range found with the other features. Finally, a backward

moving average based on five point window of data was used to smooth all features except the training lag.

A RNN was built to predict maneuver scores for one time step ahead based on features from 15 previous time steps. The RNN was composed of an input layer, one hidden LSTM layer, and an output layer. Let N be the number of maneuvers that are in the prepared data. The input layer contained $15(N + 2)$ nodes, one for each maneuver and the two created features, Training Lap and Skill, at each of the 15 time steps. The LSTM layer contained $N + 2$ nodes. The RNN transformed inputs into a vector of N outputs, one for each maneuver being predicted. The images in Figures 7 - 11 show the performance of the LSTM model for different learning durations.

Figure 7 shows the the model fit to the training data based on 100 epochs of learning. This model was trained for too few epochs, hence the poor fit to the training data.

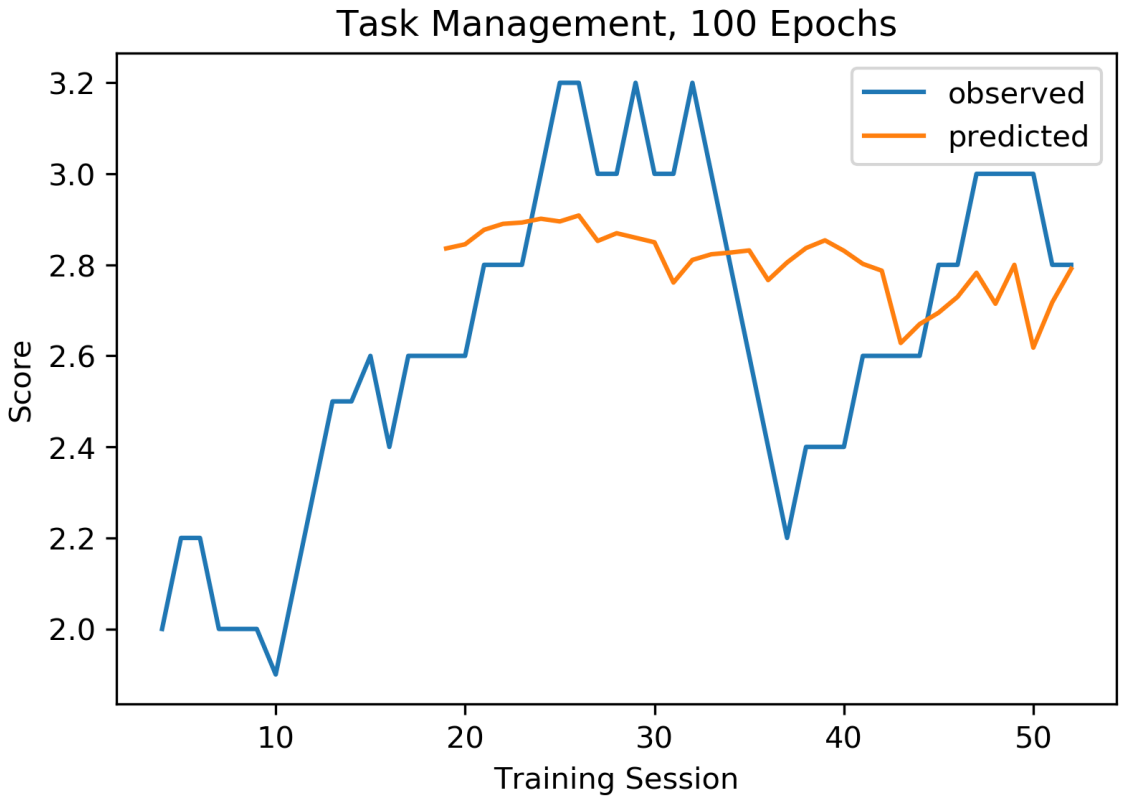

Figure 7. The model needs more learning epochs to fit the shape of the data.

Figures 8 - 9 show the fits to both the training and validation data based on 300 epochs of learning. The model has begun to fit the overall shape of the training data well. The performance for the validation data is quite poor, however. There is little correspondence between the shape or magnitude of the observed data and the model's predictions.

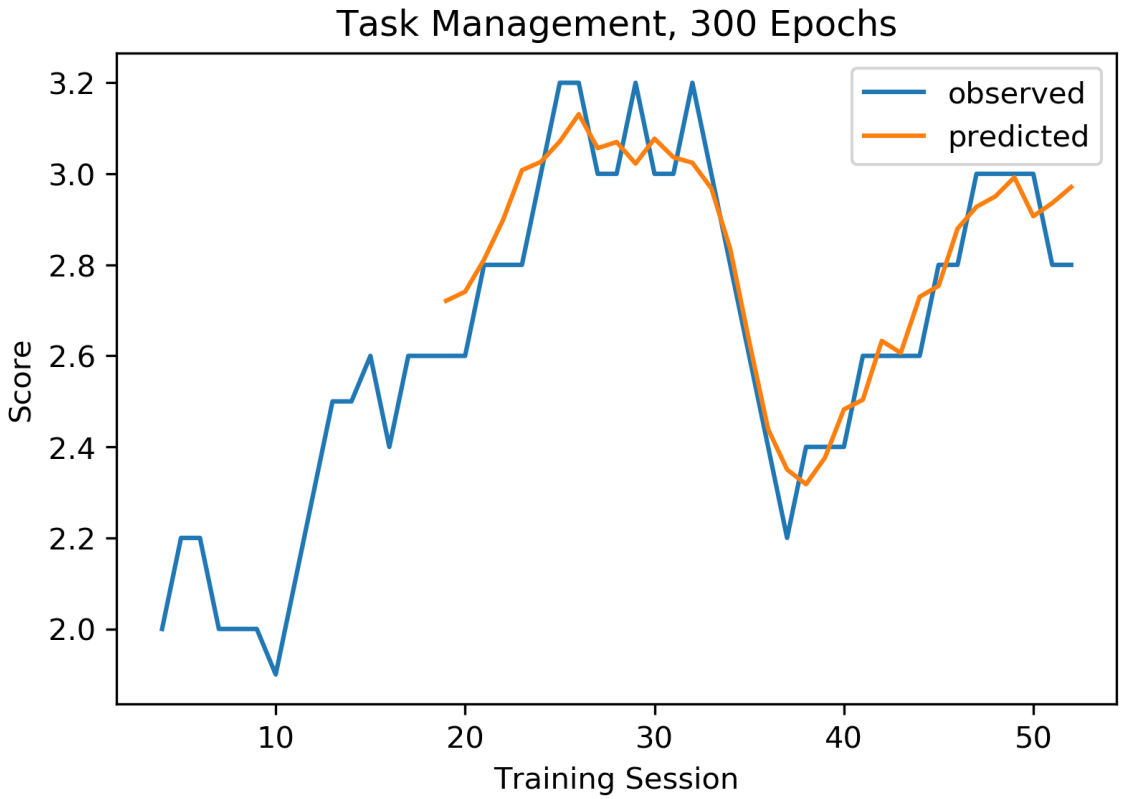

Figure 8. The model is a good fit to the training data with 300 epochs of learning.

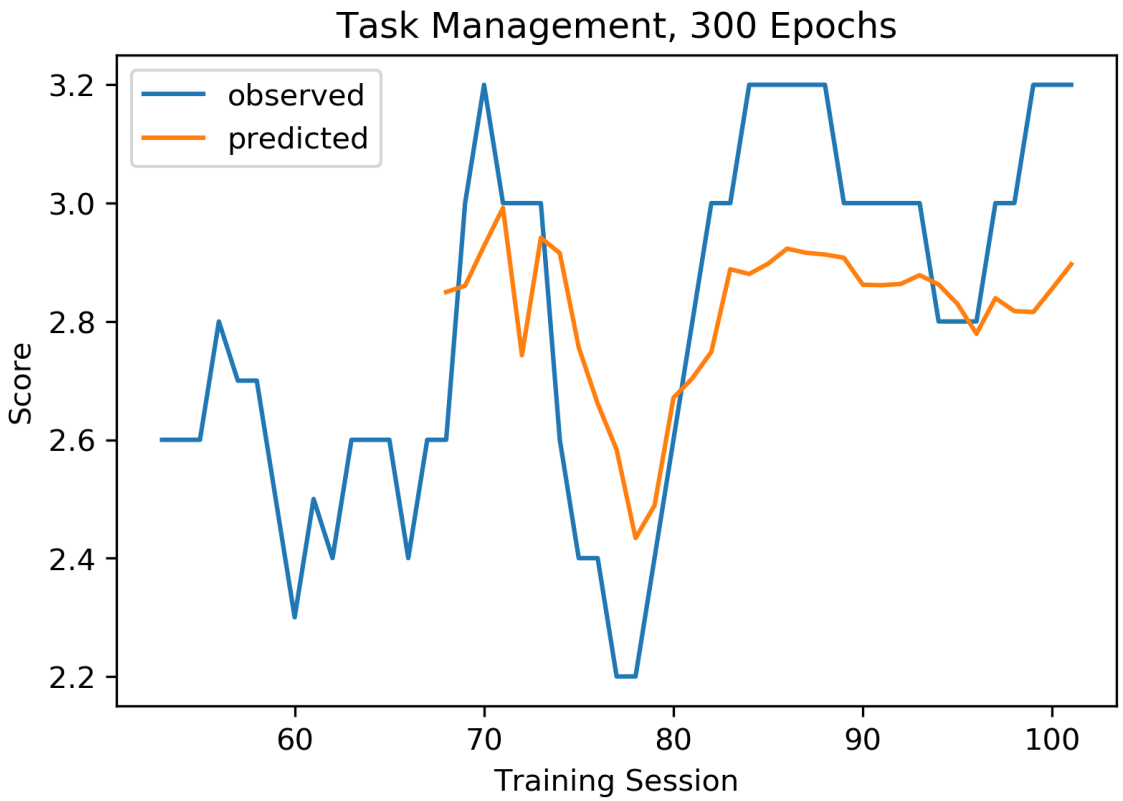

Figure 9. The model is unable to capture trends in the validation data with 300 epochs.

Figures 10 - 11 show the fits to both the training and validation data based on 500 epochs of learning. The model fits the training data well. The performance for the validation data is still quite poor.

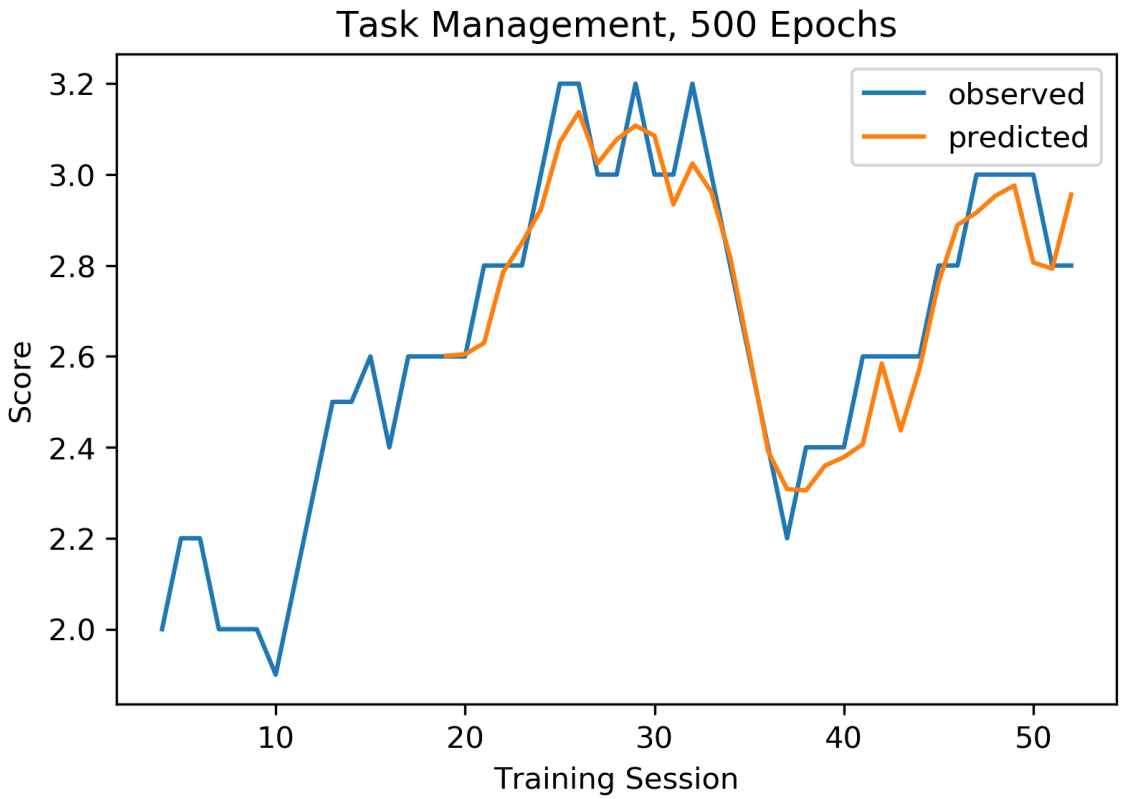

Figure 10. The model still fits to the training data with 500 epochs of learning.

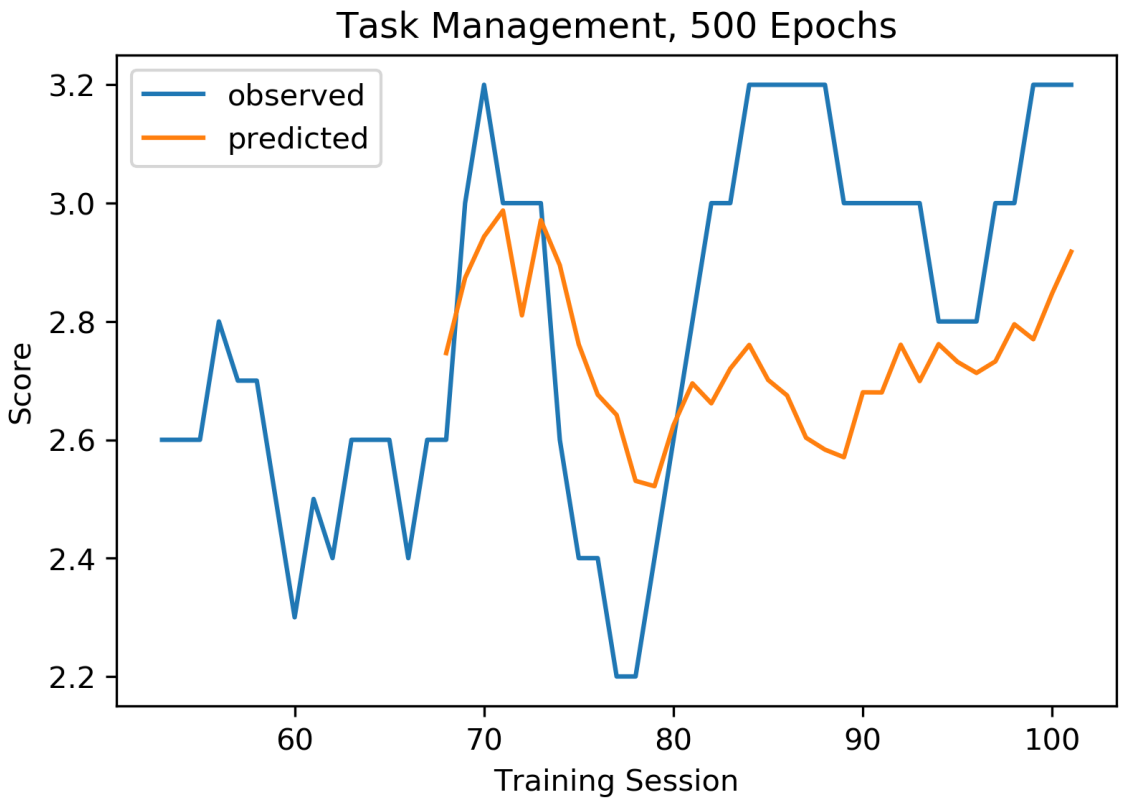

Figure 11. The fit to the validation data is not improved at 500 epochs.

The LSTM network models the training data well. However, it does not generalize well to the validation data. Table 5 shows that the mean squared error (MSE) for training predictions improves with additional learning epochs, while the validation MSE does not. This suggests that the data used is not sufficient to directly model PTN performance and that the models are overfit to the training data.

Table 5. Validation performance does not improve with more learning epochs.

Epochs	Training MSE	Validation MSE
25	0.0600	0.0533
50	0.0515	0.0624
100	0.0395	0.0572
200	0.0214	0.0689
300	0.0182	0.0731
400	0.0153	0.0765
500	0.0115	0.0718

Another RNN was built to predict binary (pass or fail) scores for each maneuver rather than the numerical scores. In addition to the data manipulation steps described previously, a new data set was created with scores below a value of three transformed to 0 (fail) and scores at three or above transformed to 1 (pass). The original data was then used to predict this new, binary data set. Predictions from this model may be viewed as the probability that a student will pass a maneuver in the following training session.

Figures 12 and 13 show fits to the binary training and validation data. The performance is quite good for the training set with peaks and troughs that correspond to the respective passes and failures in the data. However, there is no discernible pattern to the predictions for the validation data.

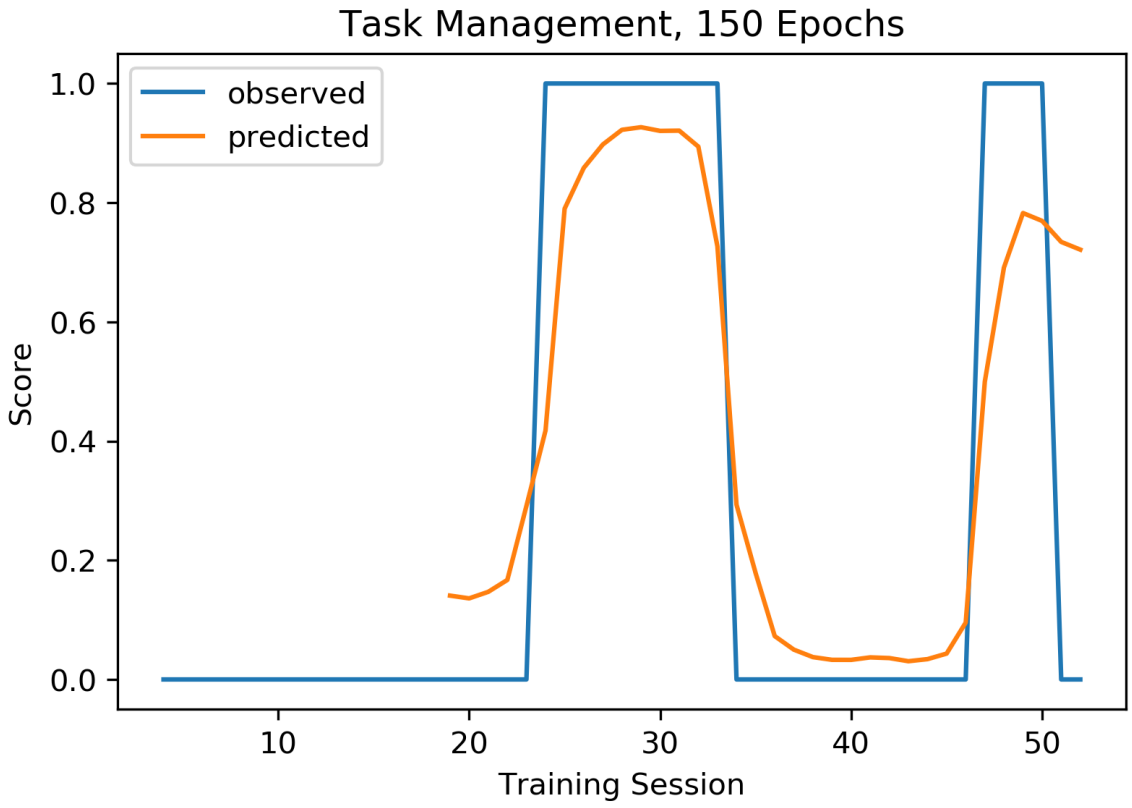

Figure 12. Modeling a binary response requires fewer learning epochs.

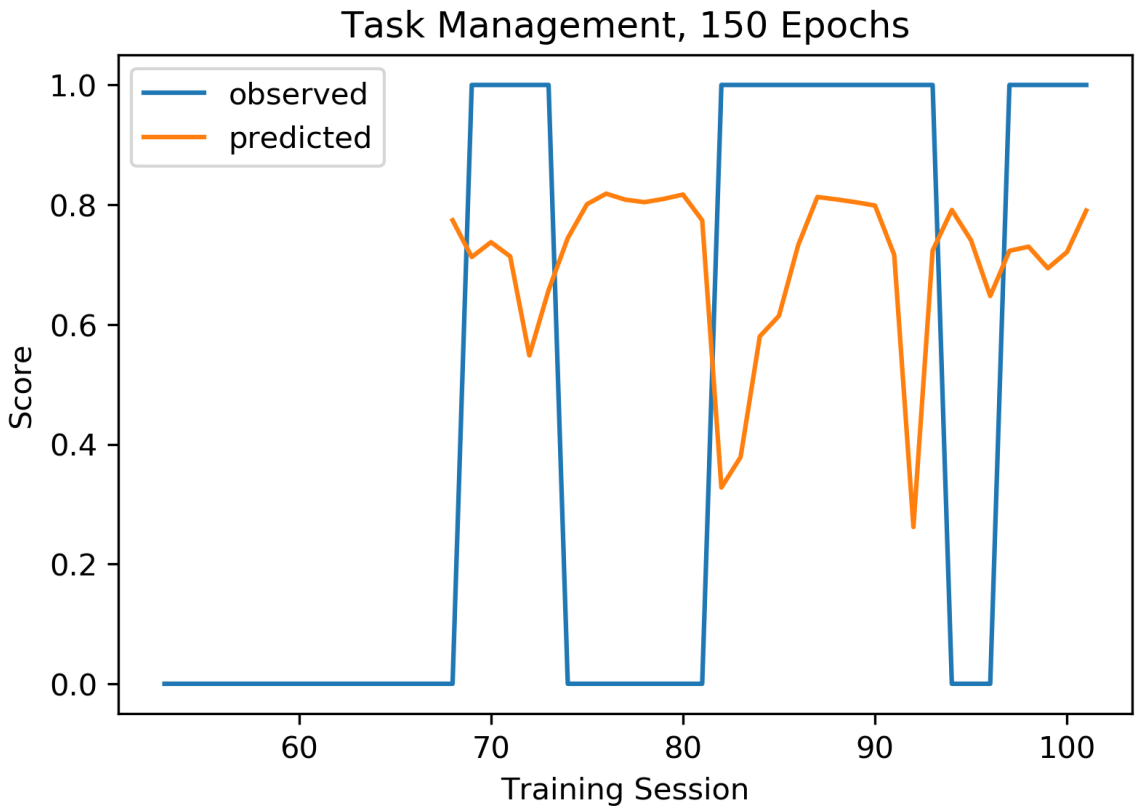

Figure 13. The binary model does not fit well to validation data.

This binary response-based RNN model suffers from similar issues to the previous RNN model. It can model the trends in the training data, but it performs poorly in validation. In particular, although there are fewer instances of failure in the validation set than in the training set, this model is still unable to predict the failures that do occur.

IV. Conclusion

Accurately predicting performance regressions among Pilot Training Next (PTN) student pilots (SPs) is a complex task due to the inherent difficult in modeling human behaviors and performance. A wide variety of physical, psychological, and emotional factors affect a pilot's flying skill. Every flight event tests a pilot's memory, situational-awareness, decision-making, workload management, and teamwork in addition to their ability to control an aircraft. However, the quantitative data collected during PTN is ambiguous and limited in scope. These issues make it difficult to model the volatile early stages of PTN when SPs are still developing competency. Improving the quality of data collected by PTN and collecting additional types of data will help enable the development of models that can more accurately predict performance regression.

PTN's current grading system does not provide insights as to why individual grades were given for training maneuvers. A student may receive a grade of Fair for a variety of different reasons. Thus, a sequence of Fairs in a particular maneuver does not necessarily indicate that a student has failed to grasp a fixed set of concepts. It is possible that students made progress in these areas while also regressing on subjects in which they previously demonstrated proficiency. Performance regressions that occur in this manner cannot be detected or predicted with the current data. Additionally, PTN's four letter grade scale is not sufficient to accurately track performance of the highly multifaceted flight maneuvers that SPs conduct. A student may show gradual improvement in all facets of a maneuver while progressing between letter grades, but this will not reflect in their numerical data until their grade suddenly jumps. Similarly, if a student lapses in performance of one or more facets of a maneuver then their score may suddenly drop. However, a simple drop in score may or may not be an actual cause for concern. These ambiguities make it difficult to model student scores

over time and to detect meaningful performance regressions at the maneuver level. A grading system with more specificity and granularity will alleviate these problems.

The PTN grading scheme can be improved by incorporating a rubric that specifically shows what the SP did well or poorly during each maneuver. Instructor Pilots (IPs) frequently leave text comments with some of this information, but there is no quantitative data provided. A rubric will provide data to link the maneuvers and more context for performance regression predictions. For example, poor performance in a set of maneuvers may be linked by a common factor such as communication skill or ability to maintain airspeed. Incorporating specific information in this manner will also increase the granularity of the data, allowing for more accurate tracking of students' skill levels for specific maneuvers. Moreover, if a set of factors that are pertinent to every maneuver can be determined, then a common rubric can be used for every graded training session, regardless of the maneuver(s) being conducted. This will allow student performance models to be constructed based on these factors rather than on individual maneuvers or groups of maneuvers. These factor based models will be faster due to using fewer variables and will not require data imputation because all factors will be observed in during each training event. The tabular form of the provided data can accommodate the additional data by simply adding one column per factor identified on the common rubric.

A psychological model may be most appropriate for predicting performance for specific flight maneuvers. Models built on psychological principles incorporate the temporal distribution of study and training events to predict performance. There may be long gaps between a SP's completion of certain maneuvers and a memory model may be used to estimate the amount of skill decay they will experience between events. A steep forgetting curve for a maneuver is indicative that a student is likely to regress in performance quickly. Psychological models models are also useful

to accelerate learning and encourage long-term knowledge retention (i.e. deter regression). Such models may be particularly useful when used in conjunction with the previously developed training recommendation system to determine the date for the recommended training event.

To conclude, the following course of action is recommended to make PTN's data more conducive to achieving its leaders' goals. Identify a set of factors that are common to most, if not all flight training events and use them to construct a grading rubric. Events from the admin and instruments categories would make a solid starting point in determining these factors. Some events may need to be specified further (e.g. basic aircraft control which may encompass several factors) or combined into a generalization (e.g. certain instruments and situational awareness). Incorporate data from simulators and any flight data recorders to the greatest extent possible. Data from flight simulators may be leveraged to quantify SP performance, giving more objectivity to evaluations by their IPs. Statistics such as deviations from expected altitude, pitch, or air speed may be useful for computing scores related to aircraft control factors. Scores for factors like checklist usage may be computed as the ratio of correctly executed items to total items executed. When computational metrics cannot be used and an ordinal score must be assessed, ensure that the levels of the grading scale are mutually exclusive and collectively exhaustive. This will ensure that scores are unambiguous. The overall score for each maneuver may then be computed as a weighted average of each factor score.

Bibliography

Betrus, A. K. (1995), 'Individualized instruction: A history of the critiques.'.

Ebbinghaus, H. (2013), 'Memory: A contribution to experimental psychology', *Annals of neurosciences* **20**(4), 155.

Ekstrand, M. D., Riedl, J. T. and Konstan, J. A. (2011), *Collaborative filtering recommender systems*, Now Publishers Inc.

Forrest, N. C. (2020), Conceptualization and Application of Deep Learning and Applied Statistics for Flight Plan Recommendation, Master's thesis, Air Force Institute of Technology.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Jastrzembski, T. S., Rogers, S. M., Gluck, K. A. and Krusmark, M. A. (2013), 'Predictive performance optimizer'. US Patent 8,568,145.

Kirk, M. (2017), *Thoughtful machine learning with python*, O'Reilly Media, Sebastopol, CA.

Koren, Y. (2008), Factorization meets the neighborhood: a multifaceted collaborative filtering model, *in* 'Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 426–434.

Koren, Y., Bell, R. and Volinsky, C. (2009), 'Matrix factorization techniques for recommender systems', *Computer* **42**(8), 30–37.

Leitner, S. (2008), *So lernt man lernen: der Weg zum Erfolg*, number 5060 *in* 'Herder-Spektrum', 16. aufl edn, Herder, Freiburg im Breisgau. OCLC: 315817348.

Manacapilli, T., O'Connell, E. and Benard, C. (2011), Customized learning: Potential air force applications, Technical report, RAND PROJECT AIR FORCE SANTA MONICA CA.

Patterson, J. and Gibson, A. (2017), *Deep learning: A practitioner's approach*, " O'Reilly Media, Inc.".

Reddy, S., Labutov, I., Banerjee, S. and Joachims, T. (2016), Unbounded human learning: Optimal scheduling for spaced repetition, *in* 'Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 1815–1824.

Robbert, A. A., Rosello, A. D., Anderegg, C. R., Ausink, J. A., Bigelow, J. H., Taylor, W. W. and Pita, J. (2015), Reducing air force fighter pilot shortages, Technical report, RAND PROJECT AIR FORCE SANTA MONICA CA SANTA MONICA.

Settles, B. and Meeder, B. (2016), A trainable spaced repetition model for language learning, *in* 'Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)', pp. 1848–1858.

Smith, B. and Linden, G. (2017), 'Two decades of recommender systems at amazon. com', *Ieee internet computing* **21**(3), 12–18.

Srivastava, A., Bala, P. K. and Kumar, B. (2020), 'New perspectives on gray sheep behavior in e-commerce recommendations', *Journal of Retailing and Consumer Services* **53**.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
25–03–2021	Master's Thesis	August 2019 — March 2021

4. TITLE AND SUBTITLE

An Examination of Potential Training Regression Recognition Algorithms for Pilot Training Next

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Gaines, Alex R., 1st Lt, USAF

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 (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-21-M-160

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 initiative to reduce the Air Force's serious pilot shortage lead to the Pilot Training Next (PTN) program. Under PTN, student pilots progress at an individual rate while making increased use of simulator-based training resources. A previous thesis used data from the first PTN class to conceptualize and prototype a student training flight scheduler. This scheduler did not consider training events required to bring students back to achieved levels of performance if in fact that student performance had regressed. This thesis examines three classes of PTN student data to determine whether student regression in training progression can be detected. A visual and two machine learning-based methods are examined and found to not predict training regression in PTN student pilots.

15. SUBJECT TERMS

pilot training next, vector auto regression, recurrent neural networks.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Raymond R. Hill, AFIT/ENS
U	U	U	UU	51	19b. TELEPHONE NUMBER *(include area code)* (937) 255-6565, x7469; rhill@afit.edu