# Enhancing Air Force Training: A Data Integration Framework

**Eric Haney, PhD**
**Lone Star Analysis**
**Dallas, Texas**
**ehaney@lone-star.com**

**Samantha Emerson, PhD & Mark Schroeder-Strong, PhD**
**Aptima**
**Woburn, Massachusetts**
**semerson@aptima.com & mschroeder@aptima.com**

## ABSTRACT

The US Air Force training enterprise is in the process of evolving to become more adaptive and responsive to the individual student. Initiatives such Air Education and Training Command's (AETC's) Pilot Training Transformation and those led by the Gaming Research Integration for Learning Laboratory (GRILL) have invested in technology and process improvements to deliver a more holistic, adaptive training environment to the warfighter. However, current data infrastructure has complicated modernization efforts. Existing training data from various sources are often stovepiped, use different data standards, and are difficult to cross-reference. Without a method to integrate data across sources, training and policy decisions cannot be driven by data.

To address this issue, this paper discusses a modular framework for synthesizing training data across disparate sources to create actionable insights. The framework must accommodate data in the state it exists today, be adaptable to change, and inform the data infrastructure of tomorrow. This paper will discuss an exploratory effort to integrate student pilot data across multiple sources and data standards into meaningful insights that can drive training. Overall, this integration across data sources will enable a more holistic view of student proficiencies to provide a basis for data-driven decisions on optimal training pathways tailored specifically to the student.

We illustrate this framework by integrating data from two primary sources (gradebook data and raw time series data logs from flight simulators) to better understand performance on specific maneuvers. First, relevant items and their associated grades are extracted from existing training gradebooks. Second, machine learning algorithms extract maneuver-level information from simulator data and quantify performance. Finally, data from both sources are used to predict performance on future maneuvers. Integration of disparate training data sources in this way offers a pathway to generate informed, data-driven decisions for Air Force training strategies.

## ABOUT THE AUTHORS

**Eric Haney** is the Chief Technology Officer at Lone Star Analysis. He is responsible for the development, deployment, and support of multiple analytics platforms, including TruNavigator™ and TruPredict™. He also leads the research and development division at Lone Star, Cipher Alchemy. As part of his work, he has been awarded two patents in edge analytics and digital twins. He holds a Ph.D. in Aerospace Engineering (University of Texas at Arlington) and a B.S. in Aerospace Engineering (Texas A&M University).

**Samantha Emerson** is a Research Scientist at Aptima, Inc. with over a decade of experience designing and executing rigorous research on human learning, thought, and language. At Aptima, her research has focused on the training, learning, and assessment, especially in United States Air Force (USAF) pilots. In collaboration with the USAF AFWERX program, she spearheaded a study examining the learning trajectories of both experienced pilots and ab initio (non) pilots as they learned to fly simulators of mature prototypes of two separate electric vertical takeoff and landing (eVTOL) vehicles. In collaboration with the USAF Air Education and Training Command (AETC) she oversaw the assessment and validation of the revised Undergraduate Pilot Training (UPT) curriculum as part of the Pilot Training Transformation (PTT) initiative. In all, her research has led to 16 scholarly publications in journals such as *Cognitive Science*, *Neuropsychologia*, and *Brain & Language* as well as 58 conference presentations including at the *Interservice/Industry Training, Simulation, and Education Conference; Vertical Flight Society*; and *CogSci*. She also serves as the secretary of SAE International's G-35 standards committee on Modelling, Simulation, and Training for Emerging Aviation Technologies and Concepts as well as the secretary for G-35's subcommittee on Flight Training and Simulation Device (FSTD) Qualification.

**Mark Schroeder-Strong** has 17 years of experience in the field of applied training effectiveness research. He is also an Associate Professor of Educational Foundations at the University of Wisconsin–Whitewater, where he teaches courses in measurement, teacher education, and development. Over the past decade, he has conducted research examining skill decay, the impact of fidelity enhancements on training effectiveness, training capability assessment techniques, and the organization and application of automated data collection for objective performance measures. Dr. Schroeder-Strong has played a significant role in the initial design and extension of the capabilities of Sim MD, a SBIR Phase II-funded technology that facilitates networked evaluations of training systems to document capabilities, identify deficiencies, and provide a path toward improvements. He was also Co-Principal Investigator on an SBIR Phase II-funded program, Predicting, Analyzing, and Tracking Training Readiness and Needs (PATTRN), which improves training programs by tracking trainee proficiencies across multiple data sources, predicting future training needs, and providing instructors with recommendations and organizational tools to deliver just-in-time training. Dr. Schroeder-Strong's academic interests lie in exploring how causal relations impact perceptions and policy in education.

# Enhancing Air Force Training: A Data Integration Framework

**Eric Haney, PhD**
**Lone Star Analysis**
**Dallas, Texas**
**ehaney@lone-star.com**

**Samantha Emerson, PhD & Mark Schroeder-Strong, PhD**
**Aptima**
**Woburn, Massachusetts**
**semerson@aptima.com & mschroeder@aptima.com**

## INTRODUCTION & BACKGROUND

The US Air Force (USAF) training enterprise is in the process of evolving to become more adaptive and responsive to the individual student[1,2]. Multiple initiatives[3,4,5] envision a future where students receive individualized, adaptive training across multiple modalities supported by the right level of fidelity in each and with training feedback immediately available to both students, instructors, and command staff. These changes all feed the larger requirement that pilots be trained in less time, using less resources, while attaining the same or higher levels of proficiency. Ongoing pilot shortages[6] and increasing cost-per-flight-hour platforms[7,8] only highlight these concerns.

The current learning ecosystem has its roots in training paradigms that can be traced to World War II-era processes. These can be characterized by 1) datasets that are captured in isolated data systems and are typically digitized versions of legacy pen-and-paper collections, 2) reliance on subject matter expertise and the subjective judgements of instructors, and 3) processes built on the assumption that all students have similar backgrounds with the same underlying level of ability and do not account for additional context regarding the student. Digital transformation of these existing systems and processes will require a fundamental re-thinking of why data is being captured and how it can be used (and re-used) to improve training outcomes.

The inertia of these existing processes, as well as the sheer scope of digital transformation in large enterprises such as the USAF, means these changes will take substantial time and necessitate incremental changeover. However, the current state of geopolitics – active conflicts in Europe and the Middle East and an elevated posture in the Indo-Pacific – requires training outcomes to improve today. Higher readiness, proficiency, and pilot production cannot wait for training data infrastructure to evolve. Innovative approaches are required that both make use of today's data systems as well as allow for growth into tomorrow's data landscape.

This paper introduces a data analytics framework that focuses on the aggregation, synthesis, and sense-making of heterogeneous student pilot training datasets. We will discuss characteristics of different existing datasets across the USAF career spectrum and how they can be standardized to provide insights about the airman's abilities and training, even in instances of sparse, noisy data. Example use cases of augmented training improvements through automated maneuver extraction, event performance assessment, and training outcome forecasting highlight potential outcomes of the framework.

[1] https://www.aetc.af.mil/News/Article-Display/Article/3631159/forging-a-foundation-basic-military-training/
[2] https://www.af.mil/News/Article-Display/Article/3833809/mission-over-function-developing-combat-effective-airmen-for-great-power-compet/
[3] https://www.af.mil/News/Article-Display/Article/2044311/pilot-training-next-begins-third-iteration-january-2020/
[4] https://www.learningprofessionals.af.mil/News/Article/2985936/upt-25-a-new-start-for-tomorrows-pilots/
[5] https://sgp.fas.org/crs/weapons/IF12257.pdf
[6] https://www.airforcetimes.com/news/your-air-force/2023/03/03/perennial-pilot-shortage-puts-air-force-in-precarious-position/
[7] https://www.defensenews.com/air/2021/12/01/with-t-7-on-the-way-why-is-acc-eyeing-a-new-trainer/
[8] https://www.gao.gov/assets/gao-23-106205.pdf

## Current State of Training Data Infrastructure & Project Motivation

Current USAF training data systems are broadly characterized by stovepipes and roadblocks. Data is often captured for a specific purpose, sometimes even on a specific data infrastructure, with little consistency or integration across other systems. This can manifest in simulators that have proprietary data standards (and lack interoperability with other systems), aircraft flight records that are never stored (and therefore cannot inform future training), or event grading records that are only applied for a subsection of a pilot's career (and therefore lack training continuity across the trajectory of the career). This landscape means any effort to describe current proficiency levels of pilots, predict future proficiency, or provide adaptive training must deal with inconsistent, missing, and unintegrated data.

Due to the urgency of improved readiness and proficiency outcomes, a path to data integration must be started today and cannot wait for long-run digital transformation initiatives to resolve flaws of the current state. The solution framework has several macro-requirements to meet the needs of the Air Force:

- *Data to Skill Mapping* – the Air Force is highly intentional about identifying which skills are important to pilot success; however, most training data are not easily correlated back to a given skill(s)
- *Data Segmentation and Extraction* – training events very often involve multiple elements that happen in sequence and / or overlap. Data analysis techniques must be able to segregate mission elements and extract them in a consistent manner without the need for human intervention.
- *Data Integration and Weighting* – standard formatting and transfer mechanisms must be in place for training records to be accessible and useful throughout a pilot's career. Multiple component measures need to be weighted according to construct coverage, recency, data quality, and transferability.
- *Predictive Validation* – analytical methods can be highly affected by missing and erroneous data. Validation mechanisms that are upstream and gatekeep data integrity can alleviate future roadblocks.
- *Updated Training Data Infrastructure* – changes will be required into how data is stored, integrated, and accessed. The need for these changes will be highlighted by roadblocks encountered during ongoing efforts

Shortcomings of USAF training have long been identified: isolated data systems, reliance on subjective assessments, and assumptions of uniform student ability. In response, various attempts have been made to address these challenges, though many have been limited in scope and impact. Efforts in areas such as automated performance assessment, predicting proficiency and performance, and developing robust training data infrastructure have shown promise. Notable research and initiatives in these areas provide a foundation for understanding the advancements and remaining challenges in USAF training.

### *Data to Skill Mapping*

Historically, the use of subjective grade sheets made data to skill mapping straightforward. Direct mappings and reliance on the judgments of subject matter experts (SMEs) to adjust for context ensured reasonable accuracy, but these methods were subject to inter-rater and intra-rater reliability issues. Additional challenges included attention span, accuracy, cognitive load, and work hours, making the process time-intensive and prone to missing critical information due to human limitations. These challenges can be categorized into three general areas:

**Data Fragmentation**: Current systems are fragmented, with data captured in isolated silos, making integration and effective utilization challenging. Frames of reference often vary within and across sites, as do data standards.
**Data Quality**: Issues such as inconsistencies, inaccuracies, and incomplete data further complicate the process of mapping data to skills.
**Human Factors**: Human limitations, including cognitive overload, fatigue, and variability in SME assessments, affect the reliability and comprehensiveness of training evaluations.

With the advent of data collection devices, we can now capture performance metrics that far exceed human capacity. For instance, in air-to-air scenarios, we can track the exact loft, pitch, and yaw angles, MACH, and altitude differential at weapons launch. We can precisely identify how many times all members of a four-ship incur Minimum Acceptable Risk (MAR) violations, how long they stay in MAR, and whether they violate specific mission elements such as acceptable levels of risk or airspace restrictions. These capabilities are beyond the scope of what a single instructor – or even a team of instructors – could achieve at the pace required for modern production goals.

This technological advancement has led to the development of new metrics that can be objectively harvested. Each metric represents a piece of a larger skill set. No single metric can fully represent a skill; instead, multiple metrics must be combined to generate a comprehensive representation of performance and proficiency. Additional data are needed to capture contexts that moderate proficiency estimates, presenting a substantial challenge.

The solution requires at least three primary elements: First, data must be integrated to provide a robust representation of knowledge, skills, and competencies. This requires mapping available objective metrics to skills and merging them with subjective holistic assessments provided by SMEs when possible. This integration aims to provide a more accurate and holistic view of a trainee's abilities. Second, a robust technological framework is required to aggregate, synthesize, and analyze the data. This framework will leverage advanced data analytics techniques to handle the vast amount of data collected. Finally, the process of combining and weighting component metrics must be transparent and comprehensible to the informed observer. This transparency ensures accountability and trust in the evaluation process.

Determining how these components come together, their weighting, and the confidence they provide in estimating trainee abilities must be addressed; and SME interpretation of the collected data and skill estimates is required. Furthermore, the way component metrics are combined and weighted cannot be opaque; it must be transparent and comprehensible to the informed observer.

### *Data Segmentation and Extraction*

Harnessing relevant metrics based on communication patterns, Time-Space-Position Information (TSPI) data, weapons logs, and relational positioning between assets and enemies at a rate of several cycles per second is a problem that has been addressed at a basic level. For example, Schreiber and Bennett (2006) demonstrated how these data points can be captured and utilized effectively to assess pilot performance on multiple process and outcome metrics almost 20 years ago. This approach of using rule-based metrics that were independently represented and combined to provide broad insight into trainee performance has been used for a subset of performance tasks: weapons employment, formation responsibilities, and mission outcomes, but has failed to extend to a broader set of skills.

The next phase is to identify, isolate, and evaluate specific behavioral and tactical events that occur within the context of a larger training activity. For instance, if a fast jet penetrates a Surface-to-Air Missile (SAM) ring, is targeted, and performs a maneuver to kinematically defeat an incoming missile, evaluating this maneuver involves more than just a binary outcome of survival. It encompasses elements like energy management, timing/reactions, and path of safety (e.g., avoiding other violations, maintaining deconfliction with other assets, or not entering another SAM ring). To accomplish this, a system needs to identify when the maneuver started, when it ended, and evaluate the contextual values that determine its success.

***Machine-Learning Approaches***. Early attempts, such as those by Zaspel (1997), used machine learning (ML) to automate performance assessment of complex skills but were limited by small, noisy datasets. More recent efforts, like those by Caballero, Gaw, Jenkins, and Johnson (2023), have captured physiological data during isolated landings in T-6 simulators. Similarly, Wilson, Scielzo, Nair, and Larson (2020) detected eye movements for takeoff, cruise, and landing assessments, and Zlatkin (2024) assessed fuel tanking performance. These successes demonstrate ML's utility in highly repetitive, controlled training contexts. However, the continuous, complex, open-loop nature of military training presents challenges such as synchronizing different data collection tools, segmenting relevant data sections, and extracting features for proficiency classification. While ML-based approaches are being developed to address these challenges, they require extensive data and focused applicability.

***Rule-Based Approaches***. Rule-based methods, prevalent before ML's rise, still offer valuable capabilities. Bessey, Waggenspack, Schreiber, and Bennett (2022) identified logic-based rules to determine key event timings in USAF training scenarios, aiding synchronization, segmentation, and performance data extraction for specific events.

Similarly, Arar and Ayan (2013) used known anchor points for performance assessment, building on McCoy and Rantanen's (2000) successes in fighter and commercial aviation domains. These techniques require knowledge of ideal training profile parameters, which are fortunately often explicitly stated in USAF training materials. For mission elements with clear parameters, such as takeoffs, landings, and one-on-one weapon engagements, rule-based algorithms may offer a resilient performance assessment approach, functioning effectively even without robust training datasets. They are labor-intensive, however, and must be reconfigured or calibrated as technology and capabilities evolve.

### *Data Integration and Weighting*

A third aspect of the solution framework involves integrating and weighting disparate data sources. In the previous 'out' maneuver example, objective network-based data collection tools can capture metrics such as range, energy loss, closure rates, airspeed, g-load, reaction time, and deconfliction spheres, which provide significant insights into the pilot's proficiency. However, these tools may not capture critical contextual information, such as acceptable levels of risk, jamming interference, special instructions from the air tasking order, operational area restrictions, and broader mission objectives. These contextual factors can significantly influence the pilot's behavior and overall performance.

SME ratings offer a holistic assessment by considering all these components but are susceptible to the pitfalls of subjective data, such as inter-rater and intra-rater reliability issues. The primary challenge with objective data lies in identifying which metrics are relevant to a given skill, determining the extent to which they cover the skill, evaluating the robustness of these measures relative to contextual variables, and deciding how they should be combined to represent the skill accurately.

When merging objective data with subjective SME ratings, it is essential to consider the weighting of each data source. This includes setting appropriate cross-validation thresholds, incorporating warnings or flags where necessary, and determining the best methods for representing the integrated data to inform training and deployment decisions effectively. By addressing these challenges, we can develop a more comprehensive and reliable framework for evaluating pilot proficiency. An effective solution must address these issues and provide a balanced, integrated representation of performance.

### *Predictive Validation*

A final aspect of the solution framework involves prediction validation, which is essential to ensure the usefulness and integrity of the data within USAF training programs. Past efforts to establish predictive models have examined some broad-brush outcomes and have been mixed. Giddings (2020) took a broad lens to potential drivers of pilot success in undergraduate pilot training – entrance exams, demographics, and initial training scores were included to provide additional context. However, results showed even the best suited ML methods (in this case Naïve Bayes) could only provide a 50% accuracy rate of predicting failed training events. Jenkins, Caballero, and Hill (2022) saw marginally better accuracy with the inclusion of additional contextual data like academic degree and source of recruitment. An internal study by the USAF from Himes, Scanland, Aleguin, and Morey (2020) demonstrated that expanding truth data to downstream outcomes after initial training at the formal training unit (FTU) could be used to highlight which upstream training elements were most likely to predict a successful or failing pilot career.

There are several takeaways from these efforts. First, data models must be adequate for the questions they are addressing. These efforts tried to make a long-term prediction from overly broad data that does not capture many of the motivational, life-event, contextual, relational, and broader contextual variables that influence relevant outcomes. Second, the expansion of truth data to downstream outcomes is essential to identify the relative value of different data sources. Third, we must consider the observed range of each predictive variable. The range of each predictive variable was significantly truncated by other selection and recruitment activities that had already removed most of the pilots that would fail, which is good to manage costs but hamstrings predictive models.

These issues represent two underlying challenges: missing data and erroneous data. Data gaps can exist for a variety of reasons. Most research on dealing with missing data deals with missing cases of data, caused by data collection, entry, or privacy issues which can be addressed through data imputation or augmentation. The more genuine problem that the current effort faces is data that is missing because it was simply not possible logistically to collect, or that its value was not recognized at the time of the event. Erroneous data is a larger problem when data generation models are inaccurate, technologies change and models that represent them are not updated, or coding errors are not detected.

The solution framework requires an extension of the data integration element here, such that SME insights should guide the specification of which variables matter and how much they matter, how variables should be combined to represent skills, and what contextual variables are needed to moderate basic representations. In this way a fuller data model can be achieved, and gaps can be identified to account for predictive limitations. Further, error detection algorithms should be implemented to ensure that auto-collected objective values that are 'out of range' are flagged as subjective ratings that seem inconsistent with objective metrics that contain performance threshold violations. This requires a data infrastructure that has several necessary components described in the next section.

### ***Training Data Infrastructure***

Future demands on training data infrastructure are substantial including: size and accessibility of data storage, integration across classified environments, data integrity, privacy, and the curation and maintenance of data sets. Yang, Yu, Lammers, and Chen (2021) provided a framework for instructor assistance through automated performance data capture and assessments; however, the assumption that ML techniques were required pushed the solution to dictate capturing of truth data that may not be financially and politically feasible for all training communities. Hernandez and colleagues (2022) focused more on the semantic layer of how training experience is described and offers a core foundation for repeatable approaches across training communities. Forrest, Hill, and Jenkins (2022) explored how such a framework can be exploited to automate the recommendation of upcoming training events to maximize learning effectiveness. These efforts point towards an unbiased data layer that is agnostic of analytical techniques is the proper approach and one that allows the most future-proof data infrastructure. There are fundamental changes needed to training data infrastructure to ensure future success are:

***Centralized Storage*** First, data must be centrally stored – efforts such as BLADE (a central repository for maintenance and logistics data) and the AFRL Data Lake (repository of detailed training logs) have begun down this path, but disconnected datasets remain, especially for detailed event data from aircraft / simulators.

***Data Standardization*** Second, data must be available in standardized data formats. This standardization may not necessarily need to be pushed to source data capture mechanisms but could instead be a translation layer that extends heterogeneous datasets into a common language. This approach of not enforcing standardization at point of data capture alleviates potential points of pushback from manufacturers, service providers, and differing communities.

***Security Flexibility*** Third, the security classification of the underlying data and derived analyses must be met head on. A holistic system touching a pilot's career of training will span from open records to unclassified to highly classified, while also incorporating PII and PHI. This cannot mean, however, that the entire system be maintained to these standards for all use cases. Instead, novel data architecture designs should be employed that segment data at the dataset (and if possible, the data field or record) level. Data anonymization, deidentification, and synthetic aggregation should also be employed to ensure broad and rapid collaboration while maintaining proper security boundaries.

***Designed for Evolution*** The last, and most philosophical, change is that design decisions need to facilitate and accommodate even more change. IT systems no longer have generational lifecycles. They are relevant for a fleeting period, and are then either rapidly improved with new functionality, or replaced. This means integrations should be loosely coupled, data properties should be generic, and storage should be agnostic of underlying data formats. Building for the future means not cementing decisions today.

To summarize, despite these various attempts to address issues in USAF training data systems, past approaches have faced significant challenges. Automated performance assessments often struggled with limited data and noisy features, while prediction efforts were hindered by broad, binary outcome measures and a lack of fine-grained performance data. Training data infrastructure efforts, meanwhile, have frequently been constrained by isolated, unstandardized data systems and the logistical difficulties of integrating data across classified environments. These shortcomings underscore the need for a holistic, integrated framework that can overcome these pitfalls. This framework must ensure consistent data capture, integration, and accessibility, while also accommodating future technological advancements. Addressing these needs requires a solution that supports the USAF's evolving training requirements through modular and multi-faceted data integration.

**Solution Framework**

Considering these previous efforts, a data integration solution must be modular and multi-faceted. Data will be used for a multitude of purposes – some of which cannot be known at the time of architecture but will instead emerge over time with new technological or algorithmic breakthroughs. The goal of the data infrastructure layer then is purely to capture data and transform it into standardized and documented formats that can be easily accessible for future efforts.

On this front, the two primary artifacts from training events – event logs and grade sheets / after-action reports – each have partial, existing compliance with a stable, reusable system.

Event logs are detailed data created by the training asset – primarily simulators and aircraft sensors. Due to the security classification and logistics of retrieving data from air vehicles, simulator data is typically more likely to be available in data systems of record although both have historically been used for teaching / de-briefing and then discarded or kept on isolated data systems. The data format of event logs is also highly variable, changing from one original equipment manufacturer to another.

Grade sheets / event reports in the USAF are primarily captured through Training Integration Management System (TIMS), Graduate Training Integration Management System (G/TIMS), and Aviation Resource Management System (ARMS) that are sequentially encountered through a pilot's career. These data systems have historically been isolated to a server hosted at each air base with limited or no syncing of data between installations. While TIMS and G/TIMS data have relative consistency in data captured, ARMS's focus on operational pilot needs means different data is capture with pilot performance being a key property excluded.

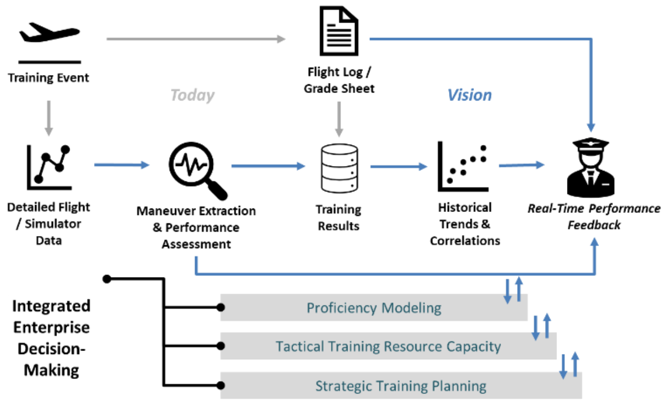

Training Event
Today
Flight Log / Grade Sheet
Vision
Detailed Flight / Simulator Data
Maneuver Extraction & Performance Assessment
Training Results
Historical Trends & Correlations
Real-Time Performance Feedback
Integrated Enterprise Decision-Making
Proficiency Modeling
Tactical Training Resource Capacity
Strategic Training Planning

**Figure 1. Data Integration Framework Topology.**

The solution architecture is shown below in Figure 1. It shows the artifacts generated by training events today and then layers on the major components required for a future of adaptive, responsive training. Training events result in both detailed flight / simulator data as well as flight log / grade sheet data. Detailed flight / simulator data often comprise millions of data points that require preprocessing such as maneuver extraction and other types of performance assessments to produce meaningful learning insights. While current sources of data tend to be siloed, it will be important in this future vision to combine both sources of data into a single repository for training results. Consolidating data in a single repository enables multi-modal analyses, correlating data from multiple sources to yield historical trends. This information can also be used to generate real-time performance feedback. Decision-making based on this model will then involve iterative feedback loops between modeling of student proficiency, available resource capacity for tactical training, and the overall strategic training plan for students. The major phases of approaching this solution state are as follows:

1. ***Consolidation and Standardization of Data*** – the consolidation side of this effort is already underway through several Digital Transformation initiatives, such as the Air Force BLADE data lake which collects maintenance and other logistics data across multiple systems (e.g. IMDS [maintenance], REMIS [reliability], CEMS [engines]) from multiple air bases. The standardization side is set to follow. For event logs, the Distributed Interactive Simulation (DIS) standard maintained by Institute of Electrical and Electronics Engineers (IEEE) has become the de facto standard for modern simulators (IEEE 1278.1-2012, 2012) and is the most likely candidate for this purpose. For event reports, the desired standards are less apparent. The most promising is the Experience

Application Programming Interface (xAPI) standard, also maintained by IEEE, which describes a learner's experiences over time (IEEE 9274.1. 1-2023, 2023). Although this phase is discussed in this paper, this is not the primary focus of the current research.

2. ***Learner Analytics*** – once data are available in an accessible and standardized format, the layering of analysis methods that extract maneuver and performance information, as well as forecast learner performance trends are required. This is the direct tie into many of the proficiency-based training and adaptive learning outcomes stated in multiple USAF training initiatives. This connection is the core of the proceeding research.
3. ***Connection to Enterprise Decision-Making*** – there exists a higher level of data-driven decision making that has not been discussed, but which is fed by the same data infrastructure and therefore shares many of the same roadblocks. Proactively planning for syllabus changes, prioritizing training technology and infrastructure investments, and future force planning for emerging requirements are all examples of decisions that could be improved when connected to live training outcomes.

## CASE STUDY

To Illustrate how this framework could be operationalized, we will apply it to real data collected during undergraduate pilot training (UPT) paired with synthetic data generated as part of the AFRL Datapalooza effort. Datapalooza was designed to generate synthetic training data for a variety of research and development purposes. These data are intended to be representative of the actual types of data that are produced throughout USAF training but at an unclassified level.

Data were produced by an experienced pilot executing the same mission of as one of three personas:  Moe (above average performance), Larry (average performance), and Curly (below average performance). A total of thirty-six data sets were produced for a simple, unclassified mission set, with twelve runs for each persona (see Figure 2). The adversaries in the data were a mix of maneuvering and non-maneuvering entities using variable maneuvers. Additionally, six of the runs each have "clutter" (non-factor entities that function as event noise in the run) while six have no clutter. The runs involve radar acquisition of an adversary, long-range missile shot, execution of a Stern Conversion, and a short-range missile shot at the adversary's close 6 o'clock.

### *Grade Sheets*

Subjective measures of performance on each of the thirty-six runs were captured in mock grade sheets (see Figure 2). The primary purpose of these grade sheets was to offer a roadmap of what made each run above average (Moe), average (Larry), or below average (Curly). Like conventional grade sheets, performance was graded on a 6-point scale that included Not Observed (NO), Unsatisfactory (0), Below Average (1), Average (2), Above Average (3), and Excellent (4). Because each persona was based on a given proficiency level, scores in the grade sheets tend to cluster around Above Average, Average, or Below Average. In practice, scores on actual grade sheets often tend to cluster around Satisfactory with little variability across items.

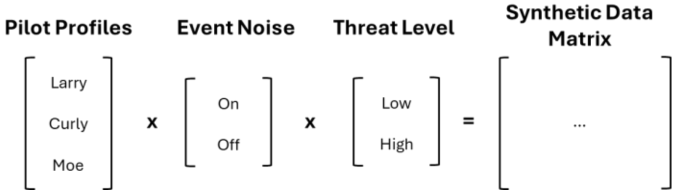
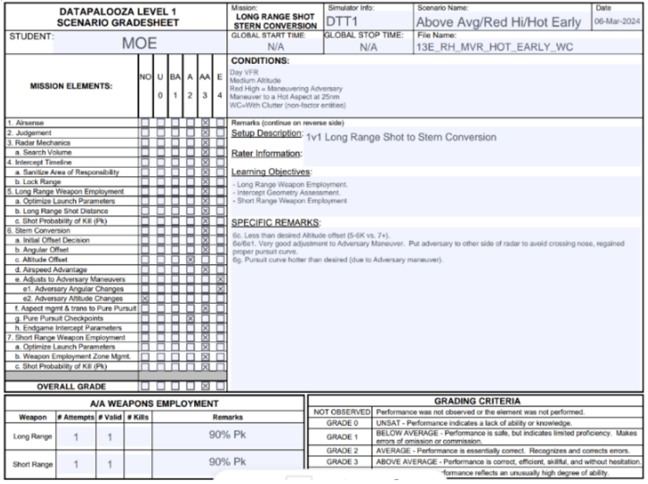


**Figure 2. Datapalooza Synthetic Dataset**



| DATAPALOOZA LEVEL 1 SCENARIO GRADESHEET | Mission: LONG RANGE SHOT STERN CONVERSION | Simulator Info: DTT1 | Scenario Name: Above Avg/Red Hi/Hot Early | Date 06-Mar-2024 |
|---|---|---|---|---|
| STUDENT: MOE | GLOBAL START TIME: N/A | GLOBAL STOP TIME: N/A | File Name: 13E_RH_MVR_HOT_EARLY_WC | |

| MISSION ELEMENTS: | NO | U<br>0 | BA<br>1 | A<br>2 | AA<br>3 | E<br>4 |
|---|---|---|---|---|---|---|
| 1. Airsense | □ | □ | □ | □ | ☒ | □ |
| 2. Judgement | □ | □ | □ | □ | ☒ | □ |
| 3. Radar Mechanics | □ | □ | □ | □ | ☒ | □ |
| a. Search Volume | □ | □ | □ | □ | ☒ | □ |
| 4. Intercept Timeline | □ | □ | □ | □ | ☒ | □ |
| a. Sanitize Area of Responsibility | □ | □ | □ | □ | ☒ | □ |
| b. Lock Range | □ | □ | □ | □ | ☒ | □ |
| 5. Long Range Weapon Employment | □ | □ | □ | □ | ☒ | □ |
| a. Optimize Launch Parameters | □ | □ | □ | □ | ☒ | □ |
| b. Long Range Shot Distance | □ | □ | □ | □ | ☒ | □ |
| c. Shot Probability of Kill (Pk) | □ | □ | □ | □ | ☒ | □ |
| 6. Stern Conversion | □ | □ | □ | □ | ☒ | □ |
| a. Initial Offset Decision | □ | □ | □ | □ | ☒ | □ |
| b. Angular Offset | □ | □ | □ | □ | ☒ | □ |
| c. Altitude Offset | □ | □ | □ | ☒ | □ | □ |
| d. Airspeed Advantage | □ | □ | □ | □ | ☒ | □ |
| e. Adjusts to Adversary Maneuvers | □ | □ | □ | □ | □ | ☒ |
| e1. Adversary Angular Changes | □ | □ | □ | □ | □ | ☒ |
| e2. Adversary Altitude Changes | ☒ | □ | □ | □ | □ | □ |
| f. Aspect mgmt & trans to Pure Pursuit | □ | □ | □ | □ | ☒ | □ |
| g. Pure Pursuit Checkpoints | □ | □ | □ | ☒ | □ | □ |
| h. Endgame Intercept Parameters | □ | □ | □ | □ | ☒ | □ |
| 7. Short Range Weapon Employment | □ | □ | □ | □ | ☒ | □ |
| a. Optimize Launch Parameters | □ | □ | □ | □ | ☒ | □ |
| b. Weapon Employment Zone Mgmt. | □ | □ | □ | □ | ☒ | □ |
| c. Shot Probability of Kill (Pk) | □ | □ | □ | □ | ☒ | □ |
| | | | | | | |
| OVERALL GRADE | □ | □ | □ | □ | ☒ | □ |

**CONDITIONS:**
Day VFR
Medium Altitude
Red High = Maneuvering Adversary
Maneuver to a Hot Aspect at 25nm
WC=With Clutter (non-factor entities)

Remarks (continue on reverse side)

Setup Description: 1v1 Long Range Shot to Stern Conversion

Rater Information:

Learning Objectives:
- Long Range Weapon Employment.
- Intercept Geometry Assessment.
- Short Range Weapon Employment

SPECIFIC REMARKS:
6c. Less than desired Altitude offset (5-6K vs. 7+).
6e/6e1. Very good adjustment to Adversary Maneuver.  Put adversary to other side of radar to avoid crossing nose, regained proper pursuit curve.
6g. Pursuit curve hotter than desired (due to Adversary maneuver).

**A/A WEAPONS EMPLOYMENT**

| Weapon | # Attempts | # Valid | # Kills | Remarks |
|---|---|---|---|---|
| Long Range | 1 | 1 | | 90% Pk |
| Short Range | 1 | 1 | | 90% Pk |

**GRADING CRITERIA**

| | |
|---|---|
| NOT OBSERVED | Performance was not observed or the element was not performed. |
| GRADE 0 | UNSAT - Performance indicates a lack of ability or knowledge. |
| GRADE 1 | BELOW AVERAGE - Performance is safe, but indicates limited proficiency.  Makes errors of omission or commission. |
| GRADE 2 | AVERAGE - Performance is essentially correct.  Recognizes and corrects errors. |
| GRADE 3 | ABOVE AVERAGE - Performance is correct, efficient, skillful, and without hesitation. |
| | formance reflects an unusually high degree of ability. |

**Figure 3. Sample Grade Sheet.**

It is important to note that the intent was not to exactly emulate USAF formats, as some elements were excluded while others were added. Non-standard items include the absence of rater (instructor pilot) information. Additionally, some grading tasks/subtasks were more detailed than a standard grade sheet, particularly the stern conversion, which was rich in numeric parameters to aid in algorithm development. Shields were maintained on entities to avoid reinitializing the simulator, and the Probability of Kill (Pk) was used to represent shot optimization, with higher Pk equating to a hit and lower Pk to a miss. Sixty percent Pk, associated with average performance, was considered a toss-up.

### *Simulator Data*

TSPI data were collected from the simulator in the comma-separated (CSV) format, with each file containing one row per time sample at a rate of 10 Hz. These rows capture state information at a moment in time for each entity within the simulation, such blue aircraft, red aircraft, and munitions. The columns in each row begin with common identifying information, including the day of the week, recording name, scenario name, participant group, research ID, vul ID, and timestamp in ISO 8601 format. The subsequent columns provide specific details for each entity, such as entity type, entity ID, result, latitude, longitude, altitude, position coordinates, airspeed, angles, heading, G-load, and various event flags.

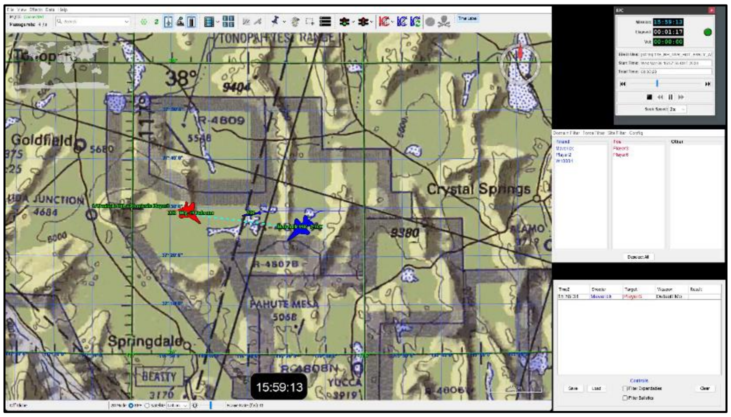


**Figure 3. Screenshot of LNCS Replay of Stern Conversion Maneuver.**

On its own, the TSPI data are not particularly useful in its raw state for understanding performance. Instead, these data need to be aggregated into meaningful performance metrics. For example, the Air Force Research Lab (AFRL) Human Performance Wing (HPW) and Aptima developed a set of tools called Performance Evaluation Tracking System (PETS) and Live, Virtual, and Constructive (LVC) Network Control Suite (LNCS) that were designed specifically for that purpose.  PETS ingests raw TSPI data and uses a rule-based approach to produce a variety of meaningful metrics, for example when and what type of munitions were deployed, numbers of hits and misses and a hit/miss ratio, time spent inside or outside a given zone, minimum and maximum ranges to key entities, and others. These metrics can be displayed on customizable dashboards or input into LNCS which provides a 3D visual to replay the events in real time (see Figure 4).

### *Graduate Training Integrated Management System (G/TIMS)*

While not part of the Datapalooza dataset, students entering a training squadron after UPT will have a record of past performance in G/TIMS. G/TIMS is a software and server platform that aggregates information across aircrew scheduling, management, training, evaluations, qualifications, resources, flight records and reporting to facilitate high-fidelity planning, proactive risk management, and decision making. It is installed at a base-level server, which then collects all training events at that location. Primary data fields include a list of training events broken down by mission element with an assigned grade from a qualified instructor at the time of the event. Grades include Not Graded (NG), Unsatisfactory (U), Fair (F), Good (G), and Excellent (E). In practice, though, grades often cluster into Satisfactory and Unsatisfactory scores, even where higher levels of attainment are technically available to the grader – most skills had 5% or less at a three or higher.

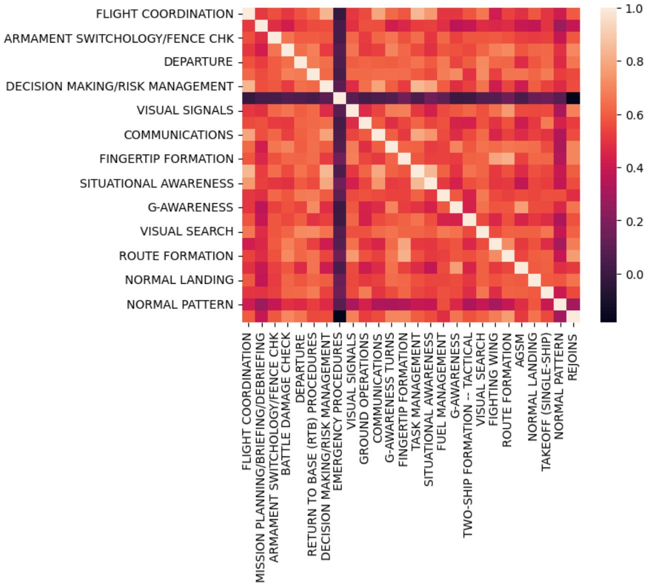


**Figure 5. Correlation Matrix of Mission Element Grades in Introduction to Fighter Fundamentals**

Furthermore, like many forms of data in the current USAF training ecosystem, each base's G/TIMS data are siloed from every other base, there is little consistency in the types of items that are graded between each base, and data are usually transferred through PDF jackets summarizing the results of training. The format of data across bases and communities is consistent, however. Thus, inclusion of G/TIMS dataset is important to the development of a data integration framework because it represents the status quo of USAF training data collection. It is the system of record for collecting and tracking training progression, and therefore is likely to remain as a fixture for any future solution. Two years of G/TIMS data was extracted from the Air Force Envision data platform for exemplary analysis.

Unsupervised clustering of historical data isolated several core skills that tend to correlate with each other and higher performance in general: flight coordination, decision-making, tasks management, and situational awareness. These results align with conclusions from previous work that showed the Inflight Planning training event from the Undergraduate Pilot Training (UPT) course was the most predictive of future failure in pilots' downstream career. The full correlation analysis for IFF skills is shown above in Figure 5.

## Application of the Framework

Applying the prescribed framework to the Datapalooza dataset provides a useful exercise. Figure 6 depicts the primary components that are discussed in the below subsections.

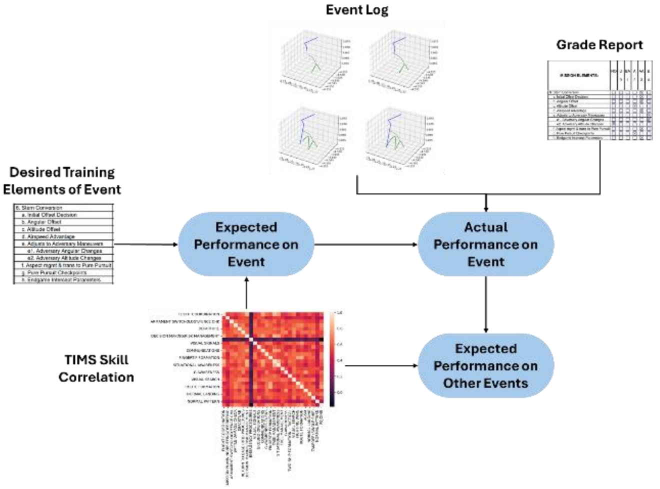

Event Log
Grade Report
Desired Training Elements of Event
6. Stern Conversion
a. Initial Offset Decision
b. Angular Offset
c. Altitude Offset
d. Airspeed Advantage
e. Adjusts to Adversary Maneuvers
e1. Adversary Angular Changes
e2. Adversary Altitude Changes
f. Aspect mgmt & trans to Pure Pursuit
g. Pure Pursuit Checkpoints
h. Endgame Intercept Parameters
Expected Performance on Event
Actual Performance on Event
TIMS Skill Correlation
Expected Performance on Other Events

**Figure 6. Framework Interactions of Stern Conversion Training Event**

### *Data to Skill Mapping & Data Segmentation & Extraction*

The application of the approach requires SMEs to deconstruct what elements are required by a particular skill, how those elements vary when it is performed poorly, adequately, or very well, and then determine what data is needed to capture this information within the training environment. In our example of a stern conversion, SMEs identified several crucial elements: The initial offset decision, the angular offset, the altitude offset, the airspeed advantage, adjustments to adversary changes in angular and altitude offsets, the management of aspect angles and transitions to pure pursuit, pure pursuit checkpoints, and endgame intercept parameters.

The next step is to determine what can be captured in the data that is available from the sim. In this case, for example, the altitude offset is easy to determine through TSPI data manipulation and can be translated into SME-defined buckets of poor, adequate, good, and excellent based on altitude differentials. This can be monitored continuously throughout the training scenario, and thresholds for each phase of the stern conversion can be specified. This allows the metric 'Altitude offset' to be generated and classified in real time. This same process can be repeated for aspect angles, intercept parameters, and adjustments to adversary maneuvers. Each metric that is created/selected can be mapped by the overall ‘stern conversion’ construct or to subconstructs as defined in the gradesheet.

Once the metrics have been mapped to their skills, the next step is to segment the data into phases of the engagement, as the optimal altitude and angular differences change as the maneuver is completed from initial positioning during the approach to weapons release. To do this manually requires a significant amount of labor, especially if the maneuver is completed as a small part of multi-ship, complex training scenario. Machine-learning algorithms are much better equipped to learn to discriminate and segment data into different phases based on large-scale changes in air speed, offset, altitude, or events such as weapons launch, or entering a weapons envelope.

In the case of a ‘stern conversion’ a simple rule-based assessment of the flight geometry was sufficient. This training element is highly prescribed by the airman community of what constitutes a well-executed maneuver from a poor one. A functional regression machine learning approach showed promise, but the small dataset limited its accuracy.

### Data Integration and Weighting

The event creates an event log which is assessed with a rule-based analysis to gauge the correctness of the maneuver geometry. At the same time, the instructor generates a grade report. At the conclusion of the event, the ‘true performance’ is then a weighted combination of the learner’s expected performance, an analytical assessment of their event log, and the expert rating of the instructor. In the case of a tightly-defined maneuver like a ‘stern conversion’ within this training context (1 vs 1 without any restrictions or SPINS) the analytical assessment of the flight logs should be weighted mostly highly – as there is little contextual variation of correct versus incorrect. Other mission types and training elements will shift this balance between the performance assessment types.

Prior to the event starting, the historical training logs of the pilot would typically be available, however the limited scope of the Datapalooza dataset eliminated their usefulness. Additionally, by assessing the correlation analysis of the baseline skills and cross-referencing that with the desired skills to be exercised during the training event, an expected performance value is created. This provides a helpful baseline for instructors, as well as creating a dampening effect for events with discrete outcomes (i.e., a success or failure on a single event informs, but does not define, a trend). The results of the individual event can then act as a Bayesian update to the trainer’s expected performance on other related tasks.

### Predictive Validation

Although the example maneuver of a `stern conversion` is highly focused and well-defined, validation of analytical approaches is still required. By comparing expected versus actual performance, outliers may be isolated and flagged for review (i.e., was the right learner’s profile entered into the system). Instructor rating deviation from actual performance may also highlight systemic issues with the algorithm and / or missing context of the event (i.e., abnormal geometry was required due to airspace requirements or verbal command was given during the event to not engage). Conversely, algorithmically generated values and violations might reveal aspects that were missed by the instructor and flagged inconsistencies would need to be resolved.

### Training Data Infrastructure

Even this limited application highlights issues in the data infrastructure layer. For instance, skill definitions at the training command level and at the operational research level are not always consistent – a translation mapping is required to align skill-to-skill correctly. Another was the continuous identification of pilots. Datapalooza datasets use anonymized IDs so training records can be correlated, however TIMS data is typically referenced by PII such as a name and social security number (SSN). Connecting the datasets requires a heightened level of information security slowing integration and limiting collaboration. This is especially deleterious when the PII is not needed for these data analyses.

## CONCLUSIONS & FUTURE WORK

This paper has explored a data integration framework for the Air Force to enhance analytically driven decision-making and improve student outcomes. Key points include the necessity of overcoming barriers between siloed datasets, leveraging multi-modal data to gain a comprehensive understanding of student performance, and using analytical assessments to reduce instructor workload while identifying areas where human expertise is crucial. Future research should focus on addressing the specific needs of individual training communities to ensure effective overlap and the assessment of emerging digital infrastructure to guarantee long-term viability.

Integrating data from various sources to inform the training of military student pilots presents significant opportunities for improvement. The barriers between siloed datasets are surmountable and should be addressed immediately to facilitate a more cohesive data environment. By utilizing multi-modal data, we can create a holistic picture of student performance, enabling more precise and effective training methods. Analytical tools can reduce the workload on instructors, allowing them to focus on areas that require their expertise, thus enhancing the overall training process. Looking forward, it is essential to assess the unique needs of different training communities to ensure a consistent and comprehensive approach. Additionally, evaluating and implementing emerging digital infrastructure will be critical to sustaining these advancements and ensuring long-term success in pilot training programs.

## ACKNOWLEDGEMENTS

This effort was funded by Air Force Research Laboratory (AFRL) Small Business Innovative Research grant AF222-D018. A special thank you to Dr. Wink Bennett (happy retirement!), Dave Malek, and Capt. John Weiland of AFRL for their oversight and guidance.

## REFERENCES

Arar, Ö. F., & Ayan, K. (2013). A flexible rule-based framework for pilot performance analysis in air combat simulation systems. *Turkish Journal of Electrical Engineering and Computer Sciences*, *21*(8), 2397-2415.

Bessey, A., Waggenspack, L., Schreiber, B., & Bennett, W. (2022). Tackling the human performance data problem: A case for standardization. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).*

Caballero, W. N., Gaw, N., Jenkins, P. R., & Johnstone, C. (2023). Toward automated instructor pilots in legacy air force systems: Physiology-based flight difficulty classification via machine learning. *Expert Systems with Applications*, *231*, 120711.

Forrest, N. C., Hill, R. R., & Jenkins, P. R. (2022). An air force pilot training recommendation system using advanced analytical methods. *INFORMS Journal on Applied Analytics*, *52*(2), 198-209.

Giddings, A. C. (2020). Predicting pilot success using machine learning. *Air Force Institute of Technology Thesis.*

Griffith, T. (2024). Leveraging AI to expand M&S toward Combined Joint All Domain Operations (CJADO). *Department of the Air Force Modeling & Simulation Summit*.

Hernandez, M., Blake-Plock, S., Owens, K., Goldberg, B., Robson, R., Center, S., & Ray, F. (2022). Enhancing the total learning architecture for experiential learning. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

Himes, T., Scanland, D., Aleguin, J., & Morey, J. (2020). 4 FW: UPT performance indicators for FTU success. *Air Combat Command (Internal).*

IEEE (2012). *Standards for Distributed Interactive Simulation—Application Protocols, Standard 1278.1-2012 (Revision of IEEE Std 1278.1-1995)*. https://ieeexplore.ieee.org/servlet/opac?punumber=6587042

IEEE (2023). *Standard for Learning Technology—JavaScript Object Notation (JSON) Data Model Format and Representational State Transfer (RESTful) Web Service for Learner Experience Data Tracking and Access, Standard 9274.1.1-2023*. https://ieeexplore.ieee.org/servlet/opac?punumber=10273183

McCoy, M. S., & Levary, R. R. (2000). A rule-based pilot performance model. *International Journal of Systems Science*, *31*(6), 713-729.

Rantanen, E. M., Talleur, D. A., Taylor, H. L., Bradshaw, G. L., Emanuel Jr, T. W., Lendrum, L., & Savoy, I. L. (2001, March). Derivation of pilot performance measures from flight data recorder information. In *11th International Symposium on Aviation Psychology*.

Jenkins, P. R., Caballero, W. N., & Hill, R. R. (2022). Predicting success in United States Air Force pilot training using machine learning techniques. *Socio-Economic Planning Sciences*, *79*, 101121.

Schreiber, B. T., & Bennett Jr, W. (2006). Distributed Mission Operations Within-Simulator Training Effectiveness Baseline Study: Summary Report. (AFRL-HE-AZ-TR-2006-0015-Vol I, 1123AS03). Air Force Research Laboratory, AZ: Warfighter Readiness Research Division.

Schreiber, B. T., Gehr, S. E., Bennett Jr, W., & LUMIR RESEARCH INST GRAYSLAKE IL. (2006). *Distributed Mission Operations Within-Simulator Training Effectiveness Baseline Study. Volume 3. Real-Time and Blind Expert Subjective Assessments of Learning* (p. 0032).

Wilson, J., Scielzo, S., Nair, S., & Larson, E. C. (2020). Automatic gaze classification for aviators: Using multi-task convolutional networks as a proxy for flight instructor observation. *International Journal of Aviation, Aeronautics, and Aerospace*, *7*(3), 7.

Yang, S., Yu, K., Lammers, T., & Chen, F. (2021, July). Artificial intelligence in pilot training and education–towards a machine learning aided instructor assistant for flight simulators. In *International Conference on Human-Computer Interaction* (pp. 581-587). Cham: Springer International Publishing.

Zaspel, J. C. (1997). Automating pilot function performance assessment using fuzzy systems and a genetic algorithm. *Oregan State University Thesis.*

Zlatkin, A. (2024). Reducing Subjectivity for Simulation-Based Pilot Training. *Department of the Air Force Modeling & Simulation Summit.*