The Method and Application of Aircrew Proficiency to High-Fidelity Mission Models in Support of Air Warfare Analysis

Roy N. Emanuel II, Alden A. Johnson, Shane A. Price, and Bryan E. Herdlick

ABSTRACT

The holistic assessment of any combat system is incomplete without evaluation of the human component. The human operator is a key, perhaps the *key, component of successful combat operations in complex environments. The Naval Air Systems Command (NAVAIR) recognized the need to consider aircrew proficiency in the achievement of warfighting objectives. In response, the Johns Hopkins University Applied Physics Laboratory (APL) developed the Proficiency-Enabled Mission Model (PEMM) to characterize the impact of operator training and readiness on mission effectiveness in the context of strike-fighter aircraft in air combat. APL's development of PEMM has advanced the state of the art for air combat modeling and simulation by introducing aircrew proficiency while executing current tactics, techniques, and procedures in the Brawler combat simulation environment. The F/A-18E/F Super Hornet defensive counter-air mission served as the initial case for proof of concept. The resulting capability informed investment decisions and training enhancements for that community. This article facilitates extension of this methodology by summarizing the process for producing a data-driven proficiency-enabled mission model with specific attention to tactics encoding, data collection, and simulation environment prerequisites.*

INTRODUCTION: THE MISSING WARFIGHTER

The US Navy transformation plan for fiscal years 2014–2016 called for an increased role of data analytics when assessing effective performance.[1] In 2014, the Naval Air Systems Command (NAVAIR) responded to the challenge with the Naval Training Analysis Framework, detailing methods for assessing aviation proficiency, cost of training, and the risk to warfighting outcomes for a given level of proficiency. Shortly after the release of these documents, the importance of US Navy training was brought to the forefront when USS *Fitzgerald* (DDG 62) and USS *John S. McCain* (DDG 56) separately collided with civilian ships, resulting in the deaths of 17 sailors. Investigators found insufficient training as a causal factor in both collisions.[2]

Training and allocation of necessary resources are also critical to success and safety in air combat, but justification for both have suffered from a lack of quantitative rigor in comparison to traditional materiel system acquisition activities. Historically, constructive mission-level models have been applied to assess combat-system performance at the hardware and software levels, with the fidelity of those system models receiving a great deal of attention. The human operators and their employment, however, have not been reflected with high fidelity and are often assumed to be flawless or acceptably represented by stochastically varied physical elements (e.g., detection range, missile trajectory) already implemented in the model(s). These assumptions are frequently invalid in the real world because operator employment varies with training, experience, system complexity, and numerous other factors.

Counter to the prevailing modeling and simulation (M&S) emphasis, the impact of aircrew training on overall system-, unit-, and force-level performance receives a lot of attention and resources from the Department of Defense (DOD), and the aviation communities in particular. Aircrews train against threat system surrogates, known as adversaries or training adversaries, to prepare for combat against combat threat aircraft. Historically, adversary aircraft have faithfully reproduced energy profiles, avionics, and weapons systems of actual threat aircraft. Today, there is growing disparity between training adversary aircraft and the combat threat aircraft they are intended to represent; foreign military capabilities have improved while adversary aircraft (e.g., F-5s, low-lot F-16s, and F-18A/Cs) used in training have not. As such, the combat outcomes resulting from training against these aging airframes are misleading (i.e., optimistic and confidence instilling) relative to the outcomes truly expected when facing peer or near-peer nations with modern fourth- and fifth-generation threat fighters. The aviation communities have a wealth of experience indicating that ineffective training can preclude aircrews from successfully employing the capabilities that give them an advantage in combat, but characterizing and quantifying that impact has proven difficult.

M&S can supplement the insights gained during training against lower-capability adversary aircraft by modeling tactics and errors and inserting fourth- and fifth-generation threat fighters into the simulation. M&S also allows for exploration of changes in tactics and system performance. Figure 1 depicts scenario clusters as proficiency, tactics, and systems change.

Tactics comprise current as well as future tactics. Current tactics are well defined and a critical part of training an operator to be effective in combat. Future tactics are typically extrapolations based on subject-matter expert (SME) knowledge of current tactics to account for changes in friendly and threat capabilities.

Modeled friendly and threat systems include training, combat, and future systems. Training systems are the current-day systems used during training events. Training systems include adversary aircraft as well as friendly aircraft capabilities, which may be restricted from use during training for safety or operational security reasons. Combat systems are combat-capable friendly and

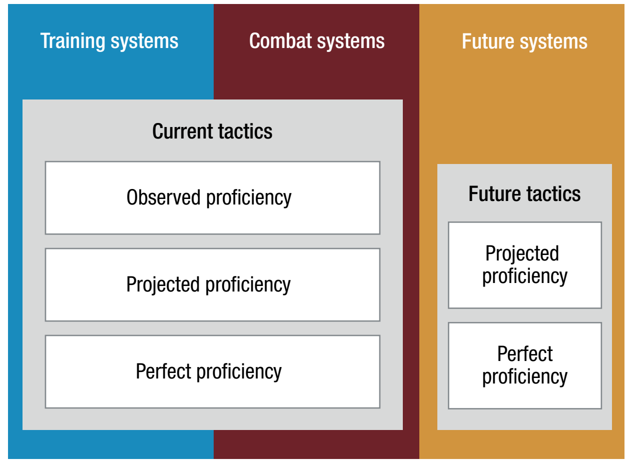


Figure 1. Scenarios of interest for supporting trade space analysis. M&S air combat scenarios vary in terms of level of aircrew proficiency, type of threat systems, and type of tactics.

threat systems. Future systems are the projected changes in friendly and threat capabilities at a common point in the future.

Proficiency values of analytical interest vary from observed aircrew proficiency during training events, to an extrapolated proficiency—possibly based on a change in training or in the complexity of Blue systems and tactics—to "perfect" proficiency [no human error relative to accepted tactics, techniques, and procedures (TTP)].

APL developed the Proficiency-Enabled Mission Model (PEMM) to address the modeling shortfalls by explicitly modeling aircrew decisions and proficiency at employing TTP. PEMM uses a growing aircrew proficiency database to elevate aircrew performance fidelity to a level comparable with that of more traditional hardware- and software-based systems in mission-level system-of-systems analyses. In the development of PEMM capabilities, the APL team extended the air combat mission modeling state of the art by adding two critical capabilities. The first is a human-readable tactics development graphical user interface (GUI). The second is the injection of error rates and magnitudes into the air combat tactics developed via the GUI. Together, these two capabilities elevate proficiency fidelity of aircrews in the Brawler[3] air-to-air combat simulation and enable extension of the proficiency and tactics models to other air combat simulations. Additionally, the methodology is extensible to other warfighting domains where TTP, metrics for execution, and operator data collection opportunities are available.

METHODOLOGY

The high-level methodology comprises inputs to a set of models that delivered results relevant to the area of study, as shown in the large boxes in Figure 2. Inputs are data collected from disparate sources such as tactics manuals, software models of platform capabilities,

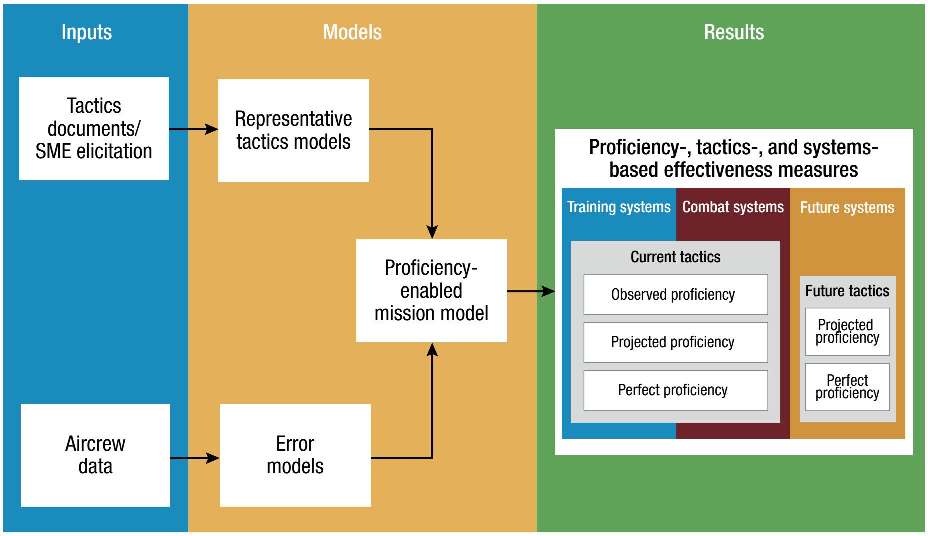


Figure 2. A high-level representation of the proficiency analysis methodology relating the inputs, models, and desired results. Inputs are data collected from disparate sources such as tactics manuals, software models of platform capabilities, and aircrew performance data collected during training events. The models use these inputs to produce results, often in the form of mission-level metrics like kill ratio (the number of enemy aircraft killed divided by the number of friendly aircraft lost). Results may inform decisions such as the assessment of training effectiveness and resource allocation or the risk a combatant commander may incur given an expected or demonstrated level of aircrew proficiency before deployment.

and aircrew performance data collected during training events. The models use these inputs to produce results, often in the form of mission-level metrics like kill ratio (the number of enemy aircraft killed divided by the number of friendly aircraft lost). Results may inform decisions such as the assessment of training effectiveness and resource allocation or the risk a combatant commander may incur given an expected or demonstrated level of aircrew proficiency before deployment.

Proficiency analyses, in this case focused on human error, use event data to find predictors for performance results. Statistical tools such as linear regression can model the relationship between the potential factors and the desired performance measures. These relationships can highlight potential factors that may improve proficiency, but they do not provide the resultant impact on a mission outcome. Additional analysis is necessary to determine whether current (or projected) proficiency achieves mission objectives. APL built PEMM to help answer this question by applying the results of a human error analysis as inputs to a mission-level simulation to explore the impacts of imperfect tactical performance on mission-level objectives.

INPUTS AND THEIR PREREQUISITES

To model warfighter proficiency with high fidelity, the analysis team must discover and/or develop critical inputs that differ significantly from traditional system-centric analyses. Inputs for a proficiency analysis can be found in manuals defining standard operating procedures and tactics, data collected from training events, and SME elicitation. In many cases, one type of data informs the other. For instance, a tactics manual defines operator tasks and the proper methods to complete them. A SME will have further insight into how the "book" tasks are actually implemented in the field. The following subsections capture such critical linkages, but an analyst attempting to apply the methodology to a new domain should be sensitive to other similar feedback opportunities that they might include in their analysis.

Tactics Documentation and SME Elicitation

A critical component of modeling warfighter proficiency is a baseline set of tasks necessary for mission success. A standard of performance for decisions and actions may be found in accepted TTP. Training agencies, known in the US Navy as Warfare Centers of Excellence (WCOEs), define standards of performance to assess and influence operator behavior toward achievement of desired outcomes that support successful mission completion within the parameters of a commander's intent. For the Naval strike-fighter community, standards of employment are published by the Naval Aviation Warfare Development Center (NAWDC) N7, also known as TOPGUN.

Employment standards at the individual and team levels are the foundational tasks and decisions necessary to model aircrew proficiency at the mission level. Capturing this information is challenging because much of the knowledge required to successfully and accurately execute tactics in the real world is in the minds of experienced operators and the institutional knowledge of training organizations. Formal tactical documentation, if it exists, is an excellent resource for the analyst, but such documentation may not have sufficient detail to model tactics as executed by the operators. Qualitative statements in tactics manuals may be difficult to represent in a model without complementary SME elicitation. For example, a comment like "must be done quickly" does not define a time frame for completion nor discuss the consequences if the task is not completed. Both are

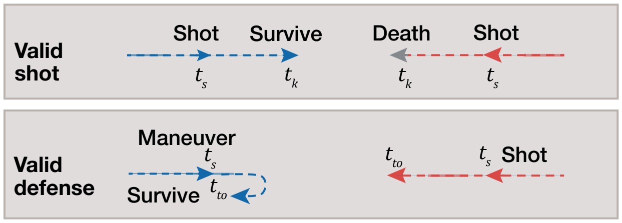


Figure 3. Unclassified example of weapon and countermeasure employment timelines. Top, Blue and Red fighters employ missiles against each other simultaneously (t_s), and the Blue fighter's weapon achieves the kill before the Red fighter's weapon (t_k). Bottom, By making a defensive maneuver when the Red fighter launched its missile (t_s), the Blue fighter defeats the missile and survives (t_{to}).

critical aspects to high-fidelity representation of aircrew proficiency in a mission model.

Two useful examples of employment standards for this discussion are missile and countermeasure employment timelines. Figure 3 shows simplified schematics of countering missile launches and using defensive maneuvers to defeat a missile. The top schematic shows Blue and Red fighters employing missiles against each other simultaneously (t_s). In this scenario and at that range, the Blue fighter's weapon achieved the kill before the Red fighter's weapon (t_k). The Blue fighter achieved tactical advantage through weapon capabilities and tactical employment decisions. The bottom example shows a defensive maneuver by the Blue fighter at the time of the Red fighter's missile launch (t_s). This action defeats the missile, resulting in the Blue fighter's survival.

The employment standards depicted in Figure 3 capture the decisions and actions that enable Blue aircrew to be both lethal and survivable. The next step in proficiency modeling is to gather data describing the types of errors that aircrews commit and the surrounding factors that contribute to these errors.

Proficiency Data Selection and Collection

NAVAIR's Training Analysis Project (TAP) has collected aircrew performance data in air combat training since 2015. Major training events, such as the Strike Fighter Advanced Readiness Program and Integrated Air Wing Training (Air Wing Fallon), provide the training scenarios and sorties for the data collection. These events are part of the Optimized Fleet Response Plan (OFRP)[4] preparing the air wing for deployment.

The data necessary for proficiency modeling fall into three general categories: performance, event, and biographical. Performance data capture the operator's actions when completing mission-essential tasks. Event data and biographical data describe the state of the training: the environment (location, weather, time of day), the operators' experience, the tactics in use, the threat presentation, and the mission objectives.

System performance and operator performance collection activities share data sources. Training range capabilities may augment these data sources by assessing the performance of the human–machine system directly. Common aviation training range capabilities include scored air-to-surface targets, surface-to-air threat training systems and emitters, and real-time missile fly-out simulations.

A data collection team derives performance metrics from the aforementioned references and standards by identifying the key areas where deviation will affect the outcome of the mission (i.e., common, critical errors). The associated data reflect operator error rates and magnitudes. Ultimately, these performance data either serve as direct inputs or provide the foundation for composite inputs to PEMM. Examples of performance data may take the form of the number of times an operator successfully performs a task or the amount of time it takes an operator to complete a task successfully. The performance data must feature a standard of performance for decisions and actions. WCOEs and other training agencies develop performance standards aligned with tactics and mission success criteria.

The analysis team found it useful to collect performance data in two general forms. The first form recorded whether an error occurred (valid) or not (invalid). For example, a valid missile launch satisfied a missile validation checklist developed by NAWDC. An invalid missile launch had one or more parameters out of tolerance as defined by the validity checklist. The second form of performance data has an error element and a magnitude element. The error element captures whether an action was executed early, on time, or late. The magnitude element captures the distance (in range or time) of the action from the acceptable value. NAWDC defined the standard for proper employment, which defined early and late and the accepted deviation from the ideal.

Event and biographical data (i.e., aircrew training and experience) are critical to describing the context of the performance data. By collecting these data, an analyst can identify which biographical markers and which event conditions have the greatest impact on aircrew performance. Event data capture the key aspects of a training event (e.g., mission type, training assets used, success criteria). These data allow an analyst to recreate the mission flow, identify critical decision points, and characterize aircrew errors. Analysis of event data can also reveal system characteristics that positively or negatively impact operator performance, including hardware and software anomalies and underused or misused system capabilities. The proficiency studies found that the qualifications and milestones the aircrew had achieved had considerable impact on performance and featured aspects such as training history, career path, and recent performance.

Table 1. Example of Data Collected from Training Events

Aircrew ID	APV	Event	Shot Validity	Kill	Defense Validity	Defense Range (nm)	Death	Attend Training
A123	A	1	Valid	Yes	—	—	—	Yes
A123	A	1	Invalid	No	—	—	—	Yes
A123	A	1	—	—	Valid	6	No	Yes
B456	B	1	—	—	Invalid	6	Yes	No
B456	B	1	Valid	Yes	—	—	—	No
C789	C	2	Valid	No	—	—	—	No

—, Not applicable.

All results from Brawler proficiency analyses are classified. What follows are unclassified data mock-ups resembling artifacts, in style only, that were presented to (and received favorably by) US Navy stakeholders. The data in Table 1 are an example of the type of data collected during a training event. Each aircrew has its own identification number to prevent attribution (Aircrew ID). Aircrew proficiency values (APVs) categorize the aircrew into similar career and training profiles. The event column captures the training event for reference to additional artifacts such as audio recordings and telemetry. The analyst records the validity and outcome of every weapon employment (shot validity) and defensive maneuver (defense validity). The kill and death columns describe the outcome of a particular engagement: Did a weapon employment result in an adversary kill? Did the defensive maneuver prevent a friendly death? Defense range is the range of an aircrew's delay in executing a defensive maneuver. And the attend training column tracks the aircrew's attendance for a particular training event.

MODELS

PEMM used three models to transform training data and tactics manuals into desired measures of effectiveness in different training and threat environments (see Figure 2). The tactics model used an APL designed-for-purpose GUI to enable rapid development, iteration, and SME review. The proficiency model aggregated the collected data and transformed it into useful inputs for the mission model. Lastly, to explore the mission trade space through stochastic variation, PEMM placed the tactics, proficiency data, and systems models into a unified constructive simulation environment based on the Brawler air combat model.

Tactics Model

DOD-accepted simulation environments such as Advanced Framework for Simulation, Integration, and Modeling (AFSIM)[5] and Brawler often use finite-state machines (FSMs), among other methods, to model operator behaviors. An FSM is an abstract, discretized machine that can be in exactly one state at any given time. The FSM transitions between states when logical criteria are satisfied. An FSM is defined by the initial state, a list of states available through transitions, and the conditions for each transition. The graphical representation of a state machine contains nodes (boxes) and edges (arrows). The nodes represent states, and the edges connecting the nodes represent allowed transitions (Figure 4).

Each "digital aircrew" in the simulation has a "mental model" of the battlespace based on inputs from sensors and communications available to the digital aircrew. Each simulation entity's worldview can be incomplete or incorrect as a result of stale, incomplete, or corrupted information. The digital operator's behavior is controlled by an FSM. For example, if an aircrew is in an ingress state and detects a threat aircraft, it could transition to an intercept state.

Another powerful capability of FSMs is nesting states within other states. This reduces repetition in the code, as all the actions of the super-state are inherited by its sub-states. As shown in Figure 4, a simulated pilot is in the defensive counter-air state and performing the tasks related for that state for the entire state-machine. FSMs

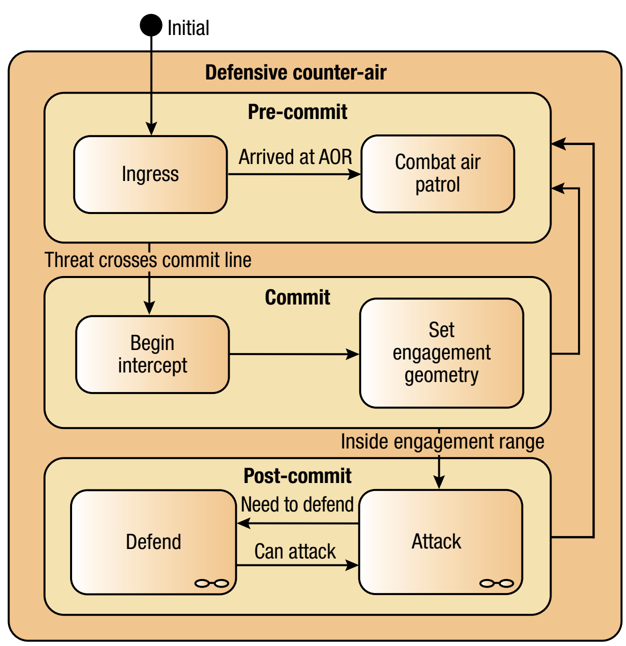


Figure 4. State-machine GUI example. The graphical representation contains nodes (boxes) and edges (arrows). The nodes represent states, and the edges connecting the nodes represent allowed transitions. AOR, area of responsibility.

offer a necessary level of control and flexibility in modeling the human decisions and behaviors in a way that supports the incorporation of error and proficiency.

Unfortunately, the FSMs defined natively in Brawler are realized by nested if-then statements written in FORTRAN. The resulting code can be hundreds of thousands of lines, and implementing, deciphering, and debugging it often requires expert-level understanding of the programming language and thorough familiarity with the particular piece of code defining the FSMs. The code complexity precludes rapid, independent SME validation of the modeled tactics.

These difficulties motivated APL to develop an FSM GUI translatable to Brawler FORTRAN. The APL team did this with independent research and development funding in conjunction with 2 Circle Consulting, Inc., which provides air warfare SMEs. In contrast to previous coding of TTP in Brawler and FORTRAN, warfighting SMEs could read and comment on tactics represented in the GUI after an hour-long training session. Furthermore, after a few short familiarization exercises, SMEs could develop and modify state-machine tactics sets without software developer supervision. This intuitive approach facilitated both collaborative development and expert validation of the PEMM through observation of the simulated fighters' behaviors.

The solution developed by APL used System Markup Language (SysML) state-machine diagrams (Figure 4) in a system architecting software (e.g., Sparx Enterprise Architect). The SysML software was then translated into a mission-level model (e.g., Brawler, AFSIM) using a translation script built in Python.

Operator Error Model

The methodology introduces aircrew proficiency into M&S by developing a statistical model of operator error. The resulting parameters from the statistical model can serve as inputs to the stochastic capabilities of mission models. Many mission models, including Brawler, support Monte Carlo analysis, which informed the outputs of the operator error model.

The errors used by PEMM manifest in two ways: frequencies and magnitudes. Error frequencies capture the probability of an error occurring at a critical point in the aircrew's tactics. Error magnitudes capture the delay, in time or range traveled, if an error occurs. Some error types, such as the probability of a valid weapon employment, only have a frequency. Other errors, such as the probability of a late defensive maneuver and the magnitude of the delay to execution, need both values. Error frequency is the probability parameter p of a Bernoulli distribution.[6] Each error opportunity causes PEMM to draw from a Bernoulli distribution with parameter p defining the probability of a valid action. The study team decided, with SME concurrence, to treat aircrew errors as independent events; p remains constant for a given aircrew throughout a mission. Study scope and schedule prevented further investigation into the utility of developing more complicated error probability models.

$$y_{i,j} \sim \text{Bernoulli}(p_j),$$

where y is the validity state of the action (valid or invalid), i is an individual observation, and j denotes the event data defining the desired group, such as the group's APV.

For an error magnitude, PEMM draws from an exponential distribution with parameter λ to assign a range from ideal execution to actual execution. λ represents a constant failure rate. The operator has an equal chance in any equal time period to recognize and correct an error (i.e., in any 3-s window, the aircrew may recognize they are late to defend and take the appropriate action). This assumption is consistent with treating errors as independent events. The formulation for the range of delay is

$$d_{i,j} \sim \lambda \exp(-\lambda t).$$

The mission model proficiency inputs for a given aircrew's APV (j) are simply the ratio of valid actions (n_{valid}) to total actions ($n_{attempts}$) for frequency errors and the average of the observed delay in time or range to execute the action (d):

$$\hat{p}_j = \frac{n_{valid,j}}{n_{attempts,j}}$$

$$\hat{\lambda}_j^{-1} = \frac{\sum_{i=1}^{n_j} d_{i,j}}{n_{attempts,j}}.$$

PEMM uses the values $\hat{p}_j$ and $\hat{\lambda}_j$ to define the parameters for its internal stochastic processes.

Another method used for the proficiency model was a Bayesian hierarchical (or multilevel) model. A rigorous discussion of this technique is beyond the scope of this article, but primers for the analysis workflow are common.[7–10] The method develops a hierarchy of relationships among the performance, biographical, and event data to produce a posterior distribution of plausible values for PEMM inputs (e.g., p, λ). Using posterior distributions as inputs for PEMM has a distinct advantage over using the single point estimators $\hat{p}$ and $\hat{\lambda}$: the posterior distribution represents uncertainty in the data and aircrew proficiency volatility from mission to mission. As an example, Figure 5 shows two different distributions of the error frequency parameter p with the same mean value ($\hat{p}$ = 0.75) but with different variance. When PEMM draws from APV A's distribution, p will rarely be less than 0.5 or greater than 0.95. In contrast, roughly 25% of APV B's distribution is less than 0.5 or greater than 0.95.

The change in variance could emulate a difference in experience level. For instance, APV A could represent

NAVAIR Public Release 2022-215 Distribution Statement A —"Approved for public release; distribution is unlimited."

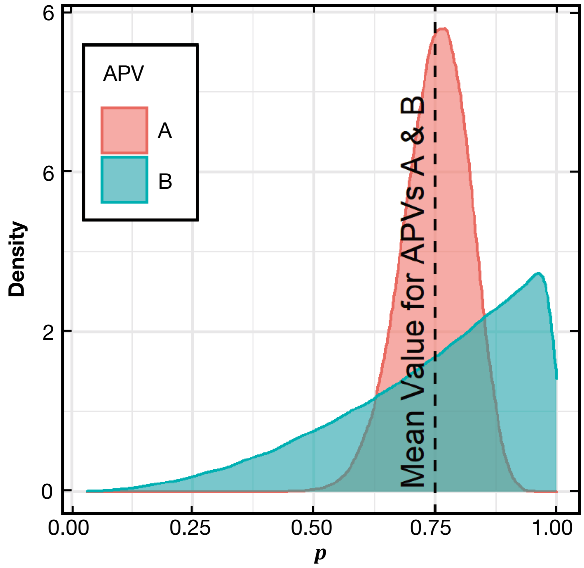


Figure 5. Demonstrating information loss when using mean values of error frequency. All APVs have the same mean error frequency, but very different density distributions.

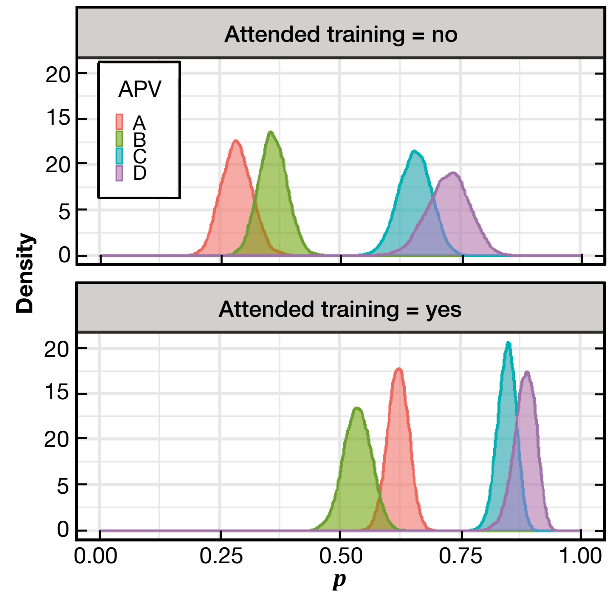


Figure 6. Example posterior distributions of *p* for different APVs with and without training. The data from Table 1 can populate a Bayesian hierarchical model to yield the posterior distributions.

an experienced group that exhibits greater consistency than APV B, a less experienced group. Because there is no expectation for the mission-level metrics to respond linearly to proficiency, capturing the differences in APV distributions will yield a richer picture of how mission-level outcomes depend on mission-to-mission aircrew performance variability.

If an analyst wants to assess the impact of a training event, the data from Table 1 can populate a Bayesian hierarchical model to yield posterior distributions for each APV with and without the training event (Figure 6). The training event improves the proficiency of all APVs by moving the bulk of each distribution toward 1 and reduces the variance of the distributions shown by narrower peaks for each APV. In this example, APV A surpasses APV B's proficiency with the training event.

PEMM: Pulling It All Together

PEMM combines the outputs of the tactics and operator error models with Brawler, a high-fidelity physics-based stochastic simulation for air combat. Brawler is government owned and is accepted for use across DOD to support air combat analyses at the engagement and mission levels. The analyst can vary the systems in the simulation, the error values, or the simulation scenario to assess the impact of training changes on mission-level metrics.

Figure 7 depicts the flow of information through the Brawler simulation. The graphical state machines are converted into Brawler-interpretable FORTRAN code with a Python programming script. The error model produces tables that Brawler can sample from the proficiency distribution for each simulation run. During

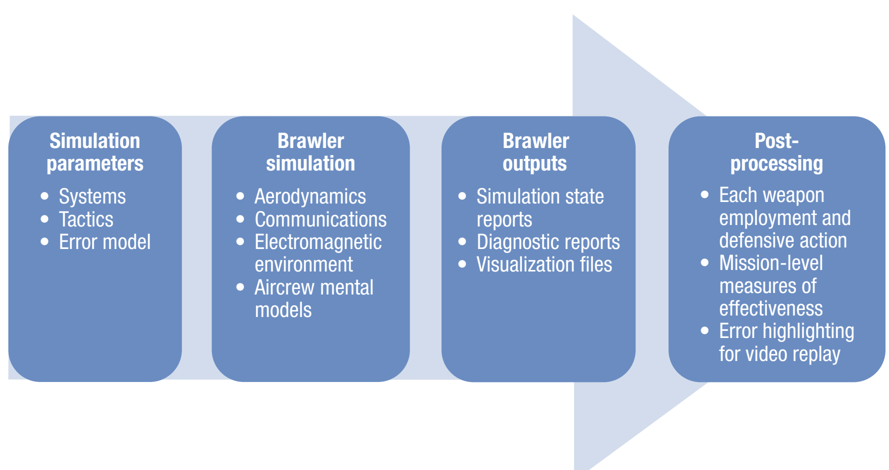


Figure 7. Schematic of information flow through the PEMM to assess the impact of proficiency on mission-level metrics. During a run, Brawler represents the aerodynamics of the aircraft and weapons, communications, aspects of the electromagnetic environment, and operators' mental models. Each run generates multiple data files recording important events and enabling video replay of the entire run. After analysts convert the data files into a format conducive to analysis, air combat SMEs review the videos of simulation runs to identify anomalous behavior and validate representative aircrew behavior.

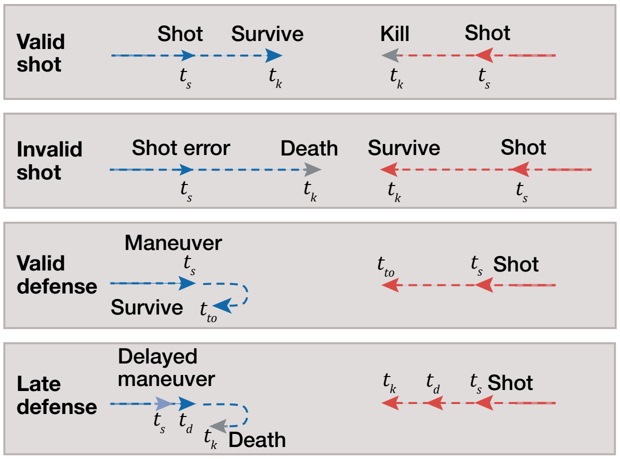


Figure 8. Comparison of missile employment and defensive maneuvers with and without aircrew errors. In the invalid shot scenario, the Blue fighter is unable to achieve proper employment parameters at the planned shot range and is subsequently killed by the Red fighter's missile after farther travel downrange. In the late defense scenario, the Blue fighter delays the defensive maneuver and is killed by the Red fighter's missile.

a run, Brawler represents the aerodynamics of the aircraft and weapons, communications, operators' mental models, and aspects of the electromagnetic environment. Each run generates multiple data files recording important events (e.g., weapon employment, aircraft death), diagnostic information to support debugging, and files to enable video replay of the entire run. Analysts convert the data files into a format conducive to analysis. Air combat SMEs review the videos of simulation runs to identify anomalous behavior and validate representative aircrew behavior. Because Brawler's reporting capability does not have a technical limit, analysts have access to and can modify the source code to extract information from the simulation.

The combination of tactics, human error, and high-fidelity modeling allowed APL to investigate the consequences of human error on the overall mission outcome—namely, if an aircrew fails to correctly employ missiles or countermeasures, how does that change its survivability and lethality for the mission? The extension of Figure 3 to include operator error is depicted in Figure 8. In the missile employment scenario, the Blue fighter is unable to achieve proper employment parameters at the planned shot range and is subsequently killed by the Red fighter's missile after farther travel downrange. In the second scenario, the Blue fighter delays the defensive maneuver and is killed by the Red fighter's missile.

Mission-level simulations such as Brawler are often much more complicated than the simple one-on-one tactic depicted in Figure 8. The scenarios often include four or more Blue aircraft and eight or more Red aircraft. The results of a single error can cascade throughout the mission because aircraft have to change their behaviors to compensate for a lost missile employment opportunity or a Blue fighter death caused by an incorrect defense.

RESULTS

Ultimately, the goals of the study direct the form of the mission model's results. Kill ratio, or loss-exchange ratio, is a common characteristic of mission success. Other informative measures include the number of friendly missiles required per enemy killed, the average range of successful shots, and the number of enemy aircraft surviving at the end of the mission (i.e., force reduction, raid annihilation, etc.). Figure 9 illustrates two ways to display mission-level metrics aggregated from a batch of Brawler simulation runs. Figure 9a shows the overall kill ratio of all Brawler simulation runs. The histogram gives an at-a-glance visual of differences between APVs. One

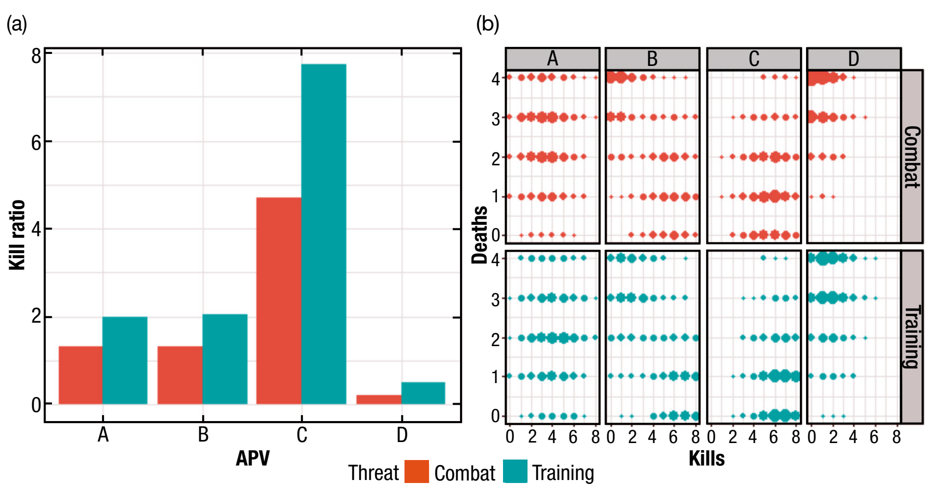


Figure 9. Kill ratio histogram example. Shown are two ways to display mission-level metrics aggregated from a batch of Brawler simulation runs. (a) The overall kill ratio of all Brawler simulation runs, giving an at-a-glance visual of differences between APVs. One drawback to this display is a loss of information when calculating the mean. (b) A heat map of kill ratio outcomes for each simulation run. The *x* axis is the number of enemy kills in a single simulation run, and the *y* axis is the number of friendly deaths in a single simulation run. The size of the circle at the kill-death vertices corresponds to the fraction of simulation runs that ended with that number of kills-deaths. Note that this version shows a difference in the distribution of individual simulation runs.

drawback to Figure 9a is a loss of information when calculating the mean. Figure 9b presents the data as a heat map of kill ratio outcomes for each simulation run. The *x* axis is the number of enemy kills in a single simulation run. The *y* axis is the number of friendly deaths in a single simulation run. The size of the circle at the kill-death vertices corresponds to the fraction of simulation runs that ended with that number of kills-deaths. Both displays show a similar overall kill ratio between APV A and APV B, but Figure 9b shows a difference in the distribution of individual simulation runs. The main mass of APV A results reside in the middle of the plot (1–2 deaths, 3–4 kills), whereas the APV B mass is at the extremes (0–1 deaths, 6–7 kills or 3–4 deaths, 1–2 kills). Both versions have value in a presentation to a stakeholder, so the analyst must consider the critical message to pass and whether to emphasize readability (Figure 9a) or information density (Figure 9b).

The context of the PEMM results depends on its inputs. Figure 10 updates the initial picture (Figure 1) to capture the areas of inquiry covered by PEMM. Block 1 uses a training threat and the proficiency observed during training events. This block answers the question, What results should we expect if we were able to repeat the event many times with the same aircrew performance? The responses could inform an instructor as to whether the results of a particular live training event were expected or were an outlier given the observed proficiency of the aircrew. Block 1 is also critical to validation and verification activities. This is the category that best

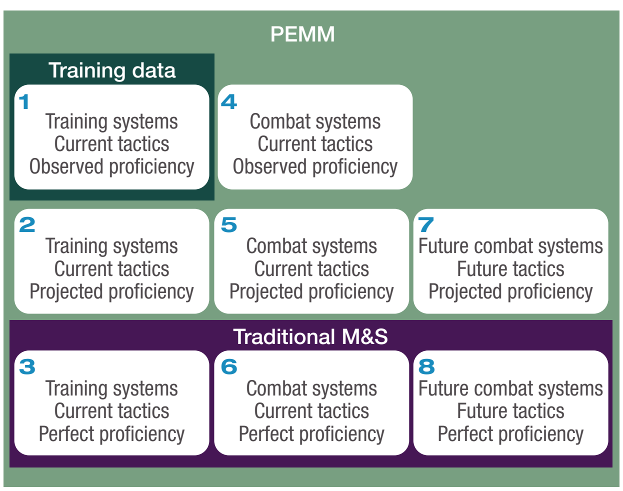


Figure 10. PEMM and traditional M&S applicability to scenarios of interest. This figure updates Figure 1 to capture the areas of inquiry covered by PEMM. Traditional M&S analysis covers the lower portion of the diagram (perfect execution), and data collected from training events cover the upper left corner of the diagram (shaded). Because of its flexibility in setting aircrew proficiency, PEMM addresses the entire M&S analysis space while being the only capability to address observed proficiency in combat and projected proficiency scenarios.

matches observations, so mission-level simulation results should align with results from the training exercises.

Block 2 keeps the training threat but modifies the proficiency to a projected value. This block answers the question, What would we expect to see in training with a different aircrew proficiency?

Block 4 changes training threat systems to combat threat systems and uses proficiency values observed in training. This block answers the question, What results should we expect if the aircrew with the observed proficiency were to face the actual threat systems rather than the training surrogate?

Block 5 uses the combat threat and a modified proficiency value. This block answers the question, What results should we expect from a projected proficiency change against the actual threat systems?

Blocks 3 and 6 establish the upper bounds for performance. The scenarios establish limitations of the systems and tactics combination with perfect execution. To improve results, the systems or the tactics must be modified. These blocks answer the question, What is the best result I can expect from improving training given the Blue systems and the training adversary or threat systems?

Blocks 7 and 8 investigate the impact of projected changes to friendly and threat system capabilities. Traditional analyses do not consider operator variance, often making an implicit assumption of perfect operator execution. The assumption of error-free operation allows the analyst to focus on the system trade space to find the necessary and sufficient capabilities to achieve victory. The results of this kind of study provide valuable inputs for defining materiel program requirements. An analysis of operator error in this scenario, currently a missed opportunity, could provide equally valuable inputs to the program's usability and training requirements.

Note that the traditional M&S analysis covers the lower portion of the diagram (perfect execution) and data collected from training events cover the upper left corner of the diagram (shaded) (current systems, current tactics, current training adversary, current proficiency). Because of its flexibility in setting aircrew proficiency, PEMM addresses the entire M&S analysis space while being the only capability to address observed proficiency in combat and projected proficiency scenarios.

LESSONS LEARNED AND IMPACT

Applying PEMM during numerous proficiency studies highlighted the critical requirement for tight integration with the data collection teams and tactical experts. In most of the studies, the data collection team and the tactics team were from the same organization (often the same people). 2 Circle Consulting, Inc., served as the data collection organization supporting the NAVAIR

TAP and provided subject-matter expertise in current F/A-18E/F air combat tactics.

The benefits of integration with the data team were twofold. First, APL could gain timely clarification on the meaning of data fields. Second, APL could provide feedback for changes to the data collected to better inform the aircrew error model and PEMM overall.

Integration with the tactics team was equally critical to iterative tactics development and refinement. APL developed an initial tactics model based on available tactics manuals using the tactics GUI. The tactics SMEs then reviewed and edited the tactics using the GUI, but, most important, they could guide the APL developers through the mindset of the aircrew while executing the tactics, thereby enhancing subsequent development of PEMM functionality. The tactics GUI represented a tremendous leap in capabilities for modeling complex tactics and enabling efficient communication between warfighting SMEs and software developers.

APL's relationship with TAP at NAVAIR has informed significant changes to the way strike-fighter aircrew train. PEMM featured prominently in a series of proficiency studies supporting NAVAIR and the Office of the Chief of Naval Operations (OPNAV) acquisition decisions in fiscal years 2017–2020. APL and TAP analyses have driven the transformation of undergraduate pilot training, the Strike Fighter Advanced Readiness Program, and Integrated Air Wing Training, also known as Air Wing Fallon.

In support of the Red Air, Training, and Proficiency for Aerial Combat (RATPAC) study (underway in 2021), APL is informing the materiel acquisitions for live-flight training adversaries, simulators, and training ranges. The NAVAIR Brawler community recognizes the tactics development GUI and error models as critical capabilities to streamline and improve future warfighting analysis for RATPAC. APL is working to transition PEMM-enabling capabilities and methodologies to NAVAIR modelers over the next 2 years. The end result of the transition is to enable modeling of future capabilities and future tactics with a projected proficiency value and use the results to inform decisions throughout the acquisition process (block 7 in Figure 10).

Each of these studies applied APL's PEMM to offer a translation of aircrew proficiency data into a representative mission-level simulation that produced measures of effectiveness used to make acquisition and training tradeoff decisions. This capability enabled analysts, operators, stakeholders, and decision-makers to develop a more intuitive understanding of potential relationships between proficiency factors and demonstrated performance.

As combat systems grow more complicated, expensive, and interconnected, program managers and analysts responsible for system design and improvement must consider the needs of the operator. Too often, analysis stops at the system level with the implicit assumption of error-free operator employment. As the United States' technological superiority diminishes, the warfighters' ability to employ their systems with winning tactics will be a critical differentiator between success and failure in the battlefield. PEMM is a novel and powerful method to show the quantitative impacts of operator proficiency to support system acquisition, training, and tactics development decisions.

ACKNOWLEDGMENTS: We acknowledge NAVAIR, OPNAV N81, and OPNAV N98 for providing the direction to conduct these studies, and in particular Dan King, Megan Roddy, Carlton Hill, and Charles Werchado. We offer profound thanks to the 2 Circle Consulting, Inc., team for their critical contributions to this field of research. Their expertise and thought leadership in the areas of air warfare and human training were an integral part of the success of this venture. Dave "Poof" Harris was a driving force behind all the proficiency studies and helped develop the methodology described here. He has been a perfect partner to APL in this endeavor. Special thanks also to Ian "Chupa" Gorski, Andrew "Schlips" Craig, Brad "Meat" Gilroy, David "FUSSY" Lash, Christopher "McFly" Stein, and Micheal "Booger" Russ, all of whom contributed directly to the development of PEMM. Thank you to everyone at APL who participated on this team in addition to the authors: Jessica Shearer, Jobin Kokkat, and Megan Lutrey. Finally, we offer posthumous gratitude to Jeff Dixon whose forethought and collaboration with Dave Harris started the independent research and development proof of concept for PEMM.

REFERENCES

[1]Department of the Navy, "Transformation plan FY 2014–2016," Washington, DC, Department of Defense, 2014, https://www.globalsecurity.org/military/library/policy/navy/usn-transformation-plan_fy2014-2016.pdf.

[2]Department of the Navy, "USS Fitzgerald, USS John S. McCain Collision Report," Washington, DC, Department of Defense, Nov. 1, 2017, https://news.usni.org/2017/11/01/uss-fitzgerald-uss-john-s-mccain-collision-report.

[3]Headquarters US Air Force/A9FM, Brawler Air Combat Simulation (Model), https://dsiac.org/models/brawler, Nov. 3, 2021.

[4]Department of the Navy, US Pacific Fleet, "Optimized fleet response plan," COMUSFLTFORCOM/COMPACFLTINST 3000.15A, Washington, DC, Department of Defense, Dec. 8, 2014.

[5] P. D. Clive, J. A Johnson, M. J. Moss, J. M. Zeh, B. M. Birkmire, and D. D. Hodson, "Advanced Framework for Simulation, Integration and Modeling (AFSIM)," in *Proc. Int. Conf. Sci. Comput*., 2015, pp. 73–77, http://worldcomp-proceedings.com/proc/p2015/CSC7058.pdf.

[6]A. O'Connor, M. Modarres, and A. Mosleh, *Probability Distributions Used in Reliability Engineering*, College Park, MD: Center for Reliability Engineering, University of Maryland, 2011, https://crr.umd.edu/sites/crr.umd.edu/files/Free%20Ebook%20Probability%20Distributions%20Used%20in%20Reliability%20Engineering.pdf.

[7]A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*, Boca Raton, FL: CRC Press, 2013.

[8]R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Boca Raton, FL: CRC Press, 2018.

[9]J. K. Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, London: Academic Press, 2014.

[10]A. Gelman, A. Vehtari, D. Simpson, C. Margossian, B. Carpenter, Y. Yao, and M. Modrák, "Bayesian workflow," arXiv, Nov. 3, 2020, https://arxiv.org/abs/2011.01808.

Roy N. Emanuel II, Space Exploration Sector, Johns Hopkins University Applied Physics Laboratory, El Segundo Field Office, El Segundo, CA

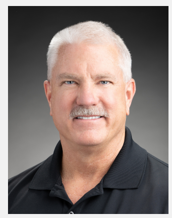
Roy N. "Snorkel" Emanuel is a project manager and section supervisor in APL's Space Exploration Sector. He has a BS in physics from the Massachusetts Institute of Technology (MIT), an MS in aeronautical and astronautical engineering from Stanford University, and a PhD in reliability engineering from the University of Maryland. During his 12 years at APL, Roy has worked on a variety of projects in the Space Exploration Sector and Force Projection Sector, including test and evaluation of space sensors and weapons systems, digital engineering environment development, resilience in acquisition and life cycle management of Navy programs, proficiency modeling of fighter pilots in air-to-air combat, and software verification and validation. Before joining APL, Roy served as a naval flight officer and accrued more than 1,000 hours and 30 combat missions in the F/A-18F Super Hornet and F-14D Tomcat. His email address is roy.emanuel@jhuapl.edu.

Alden A. Johnson, Force Projection Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Alden A. Johnson is a section supervisor in the Strike Analysis Group in APL's Force Projection Sector where he leads modeling, simulation, and analyses with a focus on air warfare applications. He has a BS in physics from the University of Massachusetts, an MS in applied mathematics from the University of Massachusetts, and an MS in aeronautics and astronautics from the University of Washington. His specialty is developing tactics models for mission-level simulations that incorporate realistic operator behavior, including the strengths and weaknesses of the aircrew. These detailed representations of the aircrew enable the human to be the central system in system-of-systems analysis. His email address is alden.johnson@jhuapl.edu.

Shane A. Price, Force Projection Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Shane A. Price is an air combat analyst in the Strike Analysis Group in APL's Force Projection Sector. He has an MS in modeling, simulation, and virtual environments from the Naval Postgraduate School. He develops air combat simulations and analyzes current and future air combat engagements to inform decision-makers and influence Department of Defense acquisitions. As a Marine Corps aviator (now retired), Shane supported the Operations Analysis Directorate of the Marine Corps Combat Development and Integration Command performing simulation analysis on a variety of high-visibility service acquisitions. His email is shane.price@jhuapl.edu.



Bryan E. Herdlick, Force Projection Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Bryan "Chum" Herdlick is program manager for naval aviation systems in APL's Force Projection Sector. He has a BS in electrical engineering from the University of Dayton, an MS in applied physics from the Naval Postgraduate School, and a PhD in systems engineering from George Washington University. He is also a graduate of the US Navy Test Pilot School and the Naval War College. His career in the military spanned operational, flight-test, and weapon-and-tactics instructor tours in the F-14 Tomcat, as well as Department of Defense acquisition tours on the Office of the Chief of Naval Operations staff and in the Navy's Air-to-Air Missiles Program Office at the Naval Air Systems Command. His current projects focus on the exploration of combat decision-making and warfighter proficiency, with applications in the design of tactical decision aids, proficiency-enabled constructive models, and extension of aircrew behavior models to crewed–uncrewed teaming initiatives. His email is bryan.herdlick@jhuapl.edu.