December 2022

Perspective
EXPERT INSIGHTS ON A TIMELY POLICY ISSUE

MARK TOUKAN, DAVID SCHULKER

Improving U.S. Air Force Training Investments

Putting Proficiency First

The U.S. Air Force (USAF) needs to make near-term investment decisions to develop training capabilities so that airmen are ready for high-end conflict in future contingencies. Yet, despite the pressing nature of such decisions, the USAF lacks the information it needs to make informed decisions for many training investments. In particular, insufficient research is available to indicate which investments are most appropriate for training needs and which should receive priority. In this Perspective, we identify five primary components of training decisions for which additional research and information are needed to support investment decisions in the training enterprise. These gaps between research and decision components, listed in Table 1, span the training enterprise.

Proficiency, the first decision component shown in the table, is particularly important for explaining elements of the other four research gaps. We highlight this component because decisions associated with proficiency measurement have a substantial downstream effect on the USAF's ability to make progress on the other four decision components. Specifically, *developing proficiency is the objective of planning*

RAND CORPORATION

Key Insights

We focused on combat pilot training to examine where the USAF requires additional information and research to support investment decisions in its operational training enterprise. The examination led to the following observations and recommendations:

- New training technologies will have only limited benefit without advancing proficiency measurement.
- New analytic approaches, for example, artificial intelligence (AI) applications, can be harnessed to improve the training enterprise but will not solve the greatest training challenges without first identifying valid performance metrics.
- The USAF should prioritize formulating performance metrics for existing competencies and experiment to identify which are most informative. The USAF should create new competencies, measurement approaches, and metrics where existing approaches do not hold up to experimentation.
- The USAF should use downtime in its emerging force generation model to generate information for metric identification, analysis, and experimentation.

for operational training and investments in operational test and training infrastructure (OTTI), and proficiency measurement is a key input to capability assessment and deployment of prescriptive analytic methods to support training.[1]

In the discussion that follows, we identify interdependencies between these research gaps and decisions and argue that efforts to address these gaps will necessarily be limited without progress on measuring and tracking proficiency. For example, the ability to select training system fidelity—how closely it replicates the flying environment—depends in part on the availability of meaningful performance measures that indicate proficiency.

The good news for the USAF is that improving proficiency measurement is possible, provided that the USAF creates the capacity for a deliberate research program to answer key questions. Further, as we describe in the conclusion, the new Air Force Force Generation (AFFORGEN) model, which allows for a phased approach to reach peak readiness (followed by recovery phases), could create the needed space for new approaches to measuring and understanding pilot learning and skill acquisition and decay.

The USAF has long recognized that a shift toward a performance-based training system is a necessity for making the most of scarce training resources and developing and presenting capabilities for mission requirements.[2] To implement such a system, the USAF requires reliable measures of performance that indicate the development and maintenance of pilot proficiency.

Abbreviations

AFFORGEN	Air Force Force Generation
AFI	Air Force Instruction
AFRL	Air Force Research Laboratory
AI	artificial intelligence
LVC	live, virtual, and constructive
OTTI	operational test and training infrastructure
PETS	Performance Evaluation Tracking System
RAP	Ready Aircrew Program
SEAD	suppression of enemy air defenses

TABLE 1

Recurrent Gaps Between Research and Decisionmaking

Key Components of Training Decisions	Research Gaps
Proficiency	• Identifying meaningful measures • Developing methods to collect measures at scale • Understanding how training tools affect proficiency
Planning for operational training	• Knowing which skills to target with training • Matching training capabilities to skills • Sequencing training to minimize skill decay • Choosing fidelity level in training systems • Determining the appropriate level of fidelity of threat models
Investments in OTTI	• Measuring cost savings • Measuring the training benefit
Capability assessment	• Measuring at the individual versus team level • Methods to aggregate metrics
Adopting prescriptive analytic methods	• Scaling subjective judgments • Automating performance assessment

As it stands, measures that flow into reporting the readiness of forces are generally based, for example, on counts of how many times a pilot has flown a sortie type or on a commander's subjective assessment of the capability of his or her squadron to perform its core tasks for a mission.[3] These measures are insufficient to tell how capable a pilot is for a mission; one pilot may need two sorties, and another may need 12 to develop a set of skills. Perhaps more important, the lack of fine-grained proficiency measures has a cascading effect on the ability of the USAF to make other investments to improve its OTTI.

The sections that follow explain each decision component, their component gaps, and interdependencies between components and gaps.

Proficiency

A recent assessment of the USAF's Ready Aircrew Program (RAP) found that the USAF lacks suitable objective metrics to assess proficiency and that there is no consensus within the combat pilot training community as to what such suitable metrics might be.[4] This finding might be expected, given that proficiency measurement is generally acknowledged as a difficult problem in the literature on pilot performance.[5] Progress on the adoption of a proficiency-based training system is hindered by (1) the challenges in specifying meaningful measures of proficiency, (2) the lack of means to collect objective metrics at scale from training events, and (3) the lack of systematic evaluations of how different training tools impact proficiency.

Gaps associated with proficiency measurement hinder progress on the four other components of training decisions. By specifying measures of proficiency, collecting measures at scale, and conducting experimentation to identify the most-informative measures, the USAF could enable progress along multiple dimensions of its training infrastructure.

Specifying Measures of Proficiency

The USAF defines proficiency in Air Force Instruction (AFI) 11-202, Volume 1, as "[a] measure of how well a task is completed. An aircrew member is considered proficient when they can perform tasks at the minimum acceptable levels of speed, accuracy, and safety."[6] Proficiency is also defined in platform-specific AFIs, with some differences across platforms.[7]

Approaches to measuring proficiency vary along two dimensions: (1) competency- versus outcome-based measures and (2) objective versus subjective measures. Table 2 provides example metrics for combinations of these approaches. Competency-based approaches identify the underlying set of "knowledge, skills, abilities, behaviors, and other characteristics" that pilots require to perform essential functions in support of executing a mission.[8] Outcome-based measurement approaches focus on the results that proficient pilots are able to achieve, such as placing a high percentage of their weapons on target.[9]

Objective measurement approaches seek to capture training data in a consistent, replicable manner. Examples of objective performance data might include statistics from actual attempts to perform combat-related tasks. Subjective measurement techniques involve expert assessment of the performance of individuals and teams.[10] For example, pilots receive qualifications after an examiner observes them performing tasks and evaluates whether they performed the tasks at an acceptable level.

Pros and Cons of Each Approach

Each approach along both dimensions has its merits and drawbacks. Outcome-based approaches most directly address the observable results of a successfully executed mission; developing competencies matters only to the extent that they help pilots achieve mission outcomes. However, outcome-based approaches are hard to generalize. Many fine-grained factors contribute to the relatively coarse outcomes that are often measured in training events, such as hit-to-kill ratios. Measuring at the level of outcome is, thus, no guarantee that the training that produced that outcome generalizes to operational scenarios.

TABLE 2
Examples of Different Proficiency Metrics

	Subjective Measures	Objective Measures
Outcome-based	Target acquired in timely fashion	Percentage of weapons on target
Competency-based	Subject-matter expert assessment of situational awareness	Eye tracking as proxy for situational awareness

Competency-based approaches hold more promise for generalization. For example, a pilot who demonstrates an ability to maintain situational awareness in a range of stressful, complex scenarios demonstrates a set of skills that may translate into scenarios that he or she has yet to encounter.

Subjective measurement approaches have the virtue of incorporating the judgment of experienced experts who may be able to identify aspects of scenarios or pilot performance that are crucial for understanding underlying proficiency but are difficult to quantify. For example, effective patterns of team communication are challenging to objectively evaluate, although an expert may be able to pinpoint specific junctions where communication can be improved to execute a mission.

On the other hand, objective measurement approaches have the virtue of being scalable and permitting comparison across individuals and over time for the same individual. Objective metrics thus permit identification of broader trends or gaps in training across the USAF.

Why Good Metrics are Hard to Come By

The discussion of pros and cons begins to suggest some reasons that it is difficult to produce quality metrics to measure and track pilot proficiency. For example, outcome-based metrics are often coarse, and many important outcomes are quite difficult—if not impossible—to simulate and assess in training contexts.[11] It is difficult to simulate the outcomes of real air combat.

Competency-based metrics are difficult to formulate because of the number and variety of skills involved in demonstrating proficiency in complex combat. For example, the Air Force Research Laboratory (AFRL) has identified mission-essential competencies for a variety of fighter platforms. One mission may involve dozens of knowledge sets and skills, making it resource intensive to formulate, measure, and track the most important metrics for a given mission.[12]

Table 3 lists examples of measurement approaches in each category that we identified in the literature. For subjective measurement, a range of approaches exist to identify and mitigate expert biases and aggregate expert assessments for greater reliability, and methods exist to calibrate expert assessments.[13] The USAF relies on subjective assessments for large parts of its standardization and evaluation program, as well as for commanders' assessments of their units.[14]

To enable more-objective approaches to measuring outcomes, the U.S. Navy has explored AI applications for the automated assessment of pilot performance in the Navy.[15] Applications in surgical simulation have shown promise in training algorithms to match expert ratings of task success to a relatively high degree.[16] Neither approach to objective measurement has yet been implemented in operational training in the USAF.

Finally, a lack of consensus on appropriate measures impedes progress on measuring combat pilot proficiency. This is a natural result of failing to rigorously demonstrate the value of a given measurement or measurement approach via experimentation. Members of the pilot training community would be more likely to adopt a measurement approach if they were confident that adopting the approach would provide value to the training enterprise.

TABLE 3

Examples of Available Measurement Approaches

	Subjective Measures	Objective Measures
Outcome-based	• Aggregation of expert assessments • Combining objective and subjective assessments • Methods to calibrate expert judgments	• Automated methods of assessing training outcomes
Competency-based	• Stan/eval program flight evaluations	• Correspondence between simulator-generated data and pilot competencies

SOURCES: For subjective measures: Brownstein et al., 2019; Ilan Yaniv, "Weighting and Trimming: Heuristics for Aggregating Judgments Under Uncertainty," *Organizational Behavior and Human Decision Processes*, Vol. 69, No. 3, 1997; and AFI 11-202, 2019. For objective measures: Brad Gilroy and Dave Harris, "Performance Assessment Using Individual Skills Linked to Mission Outcomes," presented at the Interservice/Industry Training, Simulation, and Education 2020 Conference, December 2020; Roger Smith and Danielle Julia, "Machine Learning as an Effective New Tool for Assessing Human Performance During Simulation-based Training," Interservice/Industry Training, Simulation, and Education Conference 2020, December 2020; and Brian T. Schreiber, "Transforming Training: A Perspective on the Need for and Payoffs from Common Standards," Military Psychology, Vol. 25, No. 3, 2013.

NOTE: Of the approaches listed here, stan/eval (standards and evaluations) program flight evaluations are the only ones widely used across the USAF.

Methods to Collect Measures at Scale

Identifying meaningful measures of proficiency is just one element in transitioning toward a performance-based training system. To complete this transition, the USAF would need capabilities to capture *measures of proficiency at scale*. AFRL's Performance Evaluation Tracking System (PETS) shows potential to capture data relevant to pilot competencies—such as airspeed, altitude, and distance from enemy aircraft—from simulators and training environments, although the system is not widely used.[17] Many potential data sources produced by USAF training systems exist, but there is no current requirement to capture and store data for analyzing pilot proficiency.[18] This is not primarily a research problem but, rather, requires policy changes to ensure data availability and standardization. However, solving this problem does require targeted data collection, which is not achievable without an approach for generating objective metrics.

Further development is required to identify data requirements for specific pilot skills and develop associated metrics. Experimentation would then be required to assess the predictive value of these metrics for pilot performance. Some skills may require measurement capabilities not currently included in PETS to assess, for example, pilot mental workload.

How Training Tools Affect Proficiency

The USAF would ideally know how different training tools—such as simulators, part-task trainers, or simulated environments—affect different aspects of pilot proficiency so that the appropriate tool could be chosen to target a given skill set. These tools can produce a range of objective metrics that might allow the USAF to understand impact on proficiency, but experimentation—systematically varying the training inputs so that the metrics reveal their effectiveness—is paramount to identify which measures are the most informative.[19]

To take a well-known example, the RAP requires pilots to perform certain tasks in a live setting; they can accomplish others in a simulator. Given that flying hours are sometimes scarce, it would be valuable for planners to know the proficiency benefits of live versus simulated task performance. Knowledge of how the features of the simulator contribute to proficiency could also inform future investments. Designing experiments to reveal these critical decision inputs would become possible with advances in proficiency metrics. Experiments would randomly assign pilots to different groups receiving different training offerings, and proficiency outcomes would show the relative benefits of different training modes.

Planning for Operational Training

Planning for operational training has posed a persistent challenge for the combat pilot training community, particularly in selecting the right mix of training capabilities.[20] This problem is most often posed as a mix of live and virtual sorties by mission and platform, and a number of studies have sought to address this important gap.[21] In this formulation, a mix of live and virtual sorties would be chosen to maximize proficiency while minimizing resource expenditure and the tasking of operational assets. An optimal mix, in the literal sense of the term, is unlikely to exist across all platforms and mission types; the number of factors relevant to a training mix—such as skill type, skill complexity, constraints on training resources and availability, and security concerns—implies that the answer will depend on these parameters for a specific platform and mission set. Furthermore, the answer is likely to change as these parameters change—as missions, technologies, and tactics evolve.

> Key planning decisions—such as number and frequency of training events or fidelity of training experiences—require validated proficiency metrics and the means to identify how training capabilities affect the measures. Without proficiency metrics, trainers will be unable to track how training events influence skill levels and skill maintenance, and the association with different levels of training fidelity. Trainers will also be less able to select the appropriate training tool to meet training objectives.

A key question for planners looking to make the best use of scarce resources is *how much training, and in what quantities and at what intervals, is required to produce proficient pilots?*[22] Less-proficient pilots will require more training than more-proficient pilots, but without a reliable means of measuring and tracking proficiency, the USAF has relied on standardized approaches that may under- or overshoot the amount, timing, and spacing of training events. For example, the number of sorties specified in training requirements has, at times, not been sensitive to changing training needs or the number and type of core missions for an aircraft.[23]

To help with these choices in planning operational training, research remains to be done in the following areas:

- Which skills should training target for given mission types and platforms?
- Which training events and capabilities best develop those skills?

- Relatedly, what level of fidelity is necessary to target these skills?
- At what rate do skills decay?

In most research to date, training mixes are not defined to target specific skills but rather by perceptions of the value of an event for a task or mission, such as a suppression of enemy air defenses (SEAD) mission. Matching training events and capabilities to specific skills would allow a much more targeted training program to develop proficiency. For example, one pilot may need more practice with sensor management, while another may need more practice with communication to raise proficiency for a SEAD mission.

Another challenge is identifying the right level of training fidelity and the timing of training to develop and maintain pilot proficiency.[24] Research suggests that high fidelity is not always necessary, but more work remains to be done to identify the level of fidelity necessary to develop specific types of skills and how frequently to train the skills to maintain proficiency across time.[25]

To train against specific threat scenarios, pilots require representations of adversary threat capabilities to develop proficiency. Planners must decide which new threats or variants of existing threats to replicate and the degree of fidelity with which to replicate them. The decision to invest in threat replication depends on many factors—most basically, intelligence collection—but, crucially, depends on understanding the proficiency gain of exercising against a new threat or threat variant. Without adequate measurement of proficiency, it is difficult to know whether the skills developed in training against a given generation of an adversary capability translate well into performance against the next generation of that threat. Better proficiency metrics may allow decisionmakers to prioritize investing against new threats that would not be well trained against with existing simulation capabilities.

Among the research gaps discussed here, skill decay and the fidelity of training systems and threat models have important implications for the decision gap concerning investments in OTTI. Selecting training systems that maximize skill retention can minimize unnecessary or inefficient training, and selecting the lower levels of fidelity that hit proficiency targets while achieving other training objectives prevents overinvestment in costly levels of training fidelity.

Investments in OTTI

One goal of OTTI is to enable training in operationally complex environments.[26] The vision of pilots training in operationally complex environments alongside other pilots, USAF platforms, and joint and coalition partners is both seductive and expensive. Decisionmakers may be understandably wary of diverting scarce training resources into the development of training capabilities with uncertain benefit. Credible analysis of both the costs and ben-

> The ability to estimate the cost savings and training benefit of a training investment requires proficiency measures and the ability to identify how training influences the measures. Without the ability to credibly estimate training benefits and, therefore, to specify return on investment, planners may shy away from training investments that have clear, quantifiable costs but uncertain benefits.

efits of training is necessary to make the wisest OTTI investments—whether the investments support live or virtual infrastructure—and help prioritize amongst competing resource demands.

Neither the USAF nor the other services have yet analyzed the return on investment by taking full stock of both the costs and benefits—including cost savings—of simulation-based training and live, virtual, and constructive (LVC) capabilities. This is largely the result of the difficulty in measuring benefits—especially proficiency gains—and how the benefits translate into readiness gains at higher levels of aggregation, i.e., beyond the level of the individual pilot.

While these cost estimates may be limited by data availability, e.g., training expenditures across the services, such limitations are a function of a lack of data standardization and availability rather than of fundamental research. By increasing availability of data and continuously—and consistently—measuring proficiency, the USAF would be better able to assess the return on its training investments.

Capability Assessment

The USAF reports the military readiness of its units on two broad dimensions: resources and capabilities. The former concerns a unit's resourcing and training levels for meeting operational requirements, and the latter concerns a unit's ability to perform its core tasks up to a standard. In this context, training levels are defined as the percentage of a unit that has accomplished its assigned training tasks according to some criteria, usually the tasks assigned in a RAP tasking message.[27] Of course, reporting a unit's training level in this fashion assumes that its pilots can be properly considered to be ready if they have executed a certain number of sorties for a given mission.

For the USAF to understand the readiness of its units to perform in the most complex scenarios, better proficiency metrics are required. Improved metrics could support capability assessment at both the individual and team levels.

Measuring and reporting on a unit's capability to perform its core tasks is less straightforward than reporting on the percentage of assigned tasks completed. In current practice, unit commanders subjectively assess capability at the level of the unit; however, training assessment is done at the level of the individual pilot, and units may rarely or never have the opportunity to collectively train in scenarios on which the unit must report on its capability.[28] The research challenges in this area thus involve identifying the level at which to measure capability and how to aggregate individual measures to speak to the capability of larger units.

There are many frameworks for assessing different aspects of collective performance, but none have been applied to the objective analysis of pilot team performance at scale.[29] The literatures on team cohesion, team efficacy, and team communication provide candidate markers and metrics of team performance, but existing research does not indicate which might be the most important predictors of collective performance in complex air combat or which measures could be collected at scale.[30]

Assuming that the USAF had the capability to assess collective performance objectively, it would still confront the question of whether such an approach is superior to using aggregates of individual proficiency measures as a proxy for collective proficiency. Some tasks, such as formation flying, may be candidates for assessment at the team level; however, depending on the costs and complexity of team assessment methods, it may be desirable to assess and aggregate up from the individual level. It is then an open research question as to what sets of tasks and missions are best suited for assessment at each level.

The model implicit in the USAF's aggregation practices for readiness reporting—for example, in reporting on the training level of a fighter squadron—is that individually proficient pilots aggregate up to proficient squadrons and so on up through deployed force packages.[31] This approach assumes not just that the whole is equal to the sum of its parts but also that the measures on which lower-level assessments are made are reliable.

Adopting Prescriptive Analytic Methods

The 2018 National Defense Strategy emphasizes that new technologies, which include "advanced computing, 'big data' analytics, artificial intelligence, autonomy . . . ," are essential to fighting and winning the wars of the future.[32] Further, the Air Force Chief of Staff, Gen Charles Q. Brown, has called specifically for advanced modeling, using "big data, machine learning, and AI" to be applied to understanding combat readiness of forces.[33] These examples show that leaders recognize the increasing need to support decisionmaking with *prescriptive analytics,* which is a general term for using flexible models, including machine learning techniques, to recommend courses of action to decisionmakers.

To anchor this discussion, consider three main applications of prescriptive analytics to the training and readiness domain. First, an AI system can take a subjective decision (like an instructor grading a maneuver) and create an algorithmic approximation of the scoring process that could continuously monitor the performance of the entire force. Second, an AI system could build on this performance model to forecast the performance benefits of potential training investments and could algorithmically recommend and schedule training to better optimize performance. Third, an AI system could experiment with tactics to identify new alternatives or potential vulnerabilities (this is essentially the approach involved in Defense Advanced Research Projects Agency's AlphaDogfight trials).[34]

As in the other decision components, improving proficiency measurement is a key prerequisite for moving forward with adopting each method of prescriptive analytics to inform decisions affecting capabilities. Algorithms require input data to learn how to recognize patterns that indicate good performance; that is, development of automated methods requires objective performance metrics as an input.

Automated methods hold promise for improving and scaling subjective judgments of training performance, assessing pilot skills, and forecasting pilot performance. To best harness these applications, the USAF needs to establish objective proficiency metrics to train, test, and increase trust in the underlying algorithms.

Automated assessment methods require significant investment in such areas as data interoperability, model tuning, and evaluation and safety validation. Making these investments will require understanding how they will produce improved proficiency assessment, reliability, and cost savings. Thus, objective performance metrics are important not just for implementing these methods but also for making the case for their utility as an investment.

Using AI to Assess Pilot Skills

AI may be more easily applied to assessing initial skills and making training recommendations to improve the skills. Performance metrics are more easily identified for these skills than for complex skills, which are the target of continuation training. For example, training an algorithm to identify a safely executed landing is more straightforward than training an algorithm to identify, say, communication patterns associated with a successful offensive counterair mission.[35] It is not particularly difficult to generate a large amount of relevant quantitative metrics (e.g., airspeed) to train an algorithm for such initial skills.

Using AI to Scale Subjective Judgments

AI applications that quantify and scale subjective judgments hold promise for assessing the skills and mission outcomes that are not as easily quantified. Such applications can take expert-labeled data—e.g., subjectively coded indicators of tactical or mission success—and quantify and scale that assessment when presented with new data from many training events. However, these expert judgments must accurately reflect pilot proficiency to effectively scale that assessment to assess other pilots.

The literature does not yet indicate the types of skills (along the continuum from initial skills to complex operational skills) for which automated methods are most suitable. Future work should investigate the plausibility of using inputs from aircraft sensors or simulators to assess performance and identify the characteristics of skills that are most amenable to automated assessment. Some factors relevant to individual task performance, such as team context, may be more easily spotted by experts, while automation may be appropriate to account for factors at the level of the pilot, aircraft, or simulation scenario that are more easily quantifiable. Future work could also investigate how to best leverage expert input alongside automated assessment methods.

Recommendations

In building out its training infrastructure to prepare for future conflicts, the USAF should prioritize measuring proficiency. Each of the other investment areas discussed here hinges in some important respect on the development of proficiency measurement and associated performance metrics. The problem the USAF then faces is what immediate steps to take toward measuring proficiency across the pilot training enterprise.

What Should the Air Force Do in the Short and Longer Term?

In the short term, the USAF should *prioritize formulating metrics for existing competencies and conducting experi-*

ments to identify which are most predictive of pilot performance. As mentioned previously, the USAF has already gone partway down this path with AFRL's work on competencies and the development of capabilities to pull information from training systems and environments. What is needed now is a *systematic approach to refine and carry forward this work and to develop new competencies, measurement capabilities, and performance metrics where existing approaches do not hold up to experimentation*.

Time and resources are scarce for meeting training requirements as they currently exist. The forthcoming AFFORGEN model suggests how the USAF may accommodate a campaign of experimentation and metric identification. In this new model, pilots will go through four six-month phases in a 24-month cycle. In the first phase, pilots will reset from deployments or taskings. In subsequent phases, pilots will prepare for new deployments with increasingly complex training events and exercises.[36] The first phase is intended to allow time for the USAF to balance modernization and long-term readiness priorities with meeting the short-term needs of combatant commanders. *The USAF could use this downtime to generate information for metric identification, analysis, and experimentation*. With AFFORGEN still in development across the major commands, *now* is the time for the USAF to decide what kinds of activities—including making progress toward proficiency measurement—could and should be done in each phase of the cycle.

For example, the number and types of virtual sorties that may be required to generate sufficient data to assess pilot performance in a SEAD mission may exceed what RAP requires and may therefore reduce the number of sorties available for other mission sets. This approach would thus come at the cost of allowing a temporary slippage in training for immediate needs—a risk to current readiness. However, the risk of forgoing these activities, which would enable the assessment of readiness for high-end conflict and longer-term readiness goals, is even greater.

What the Air Force Should Not Do

Given the range of efforts and technologies that have been thrown at the problem of measuring proficiency, it is important to identify approaches that are *unlikely to prove helpful*. As we discussed, there is great promise in using AI applications to enhance operational training, but *AI alone will not solve this problem*. Investing in automated methods without first developing validated performance metrics is putting the cart before the horse; algorithms need to be trained to recognize proficiency in the data that are presented to them. Because trust in automation is paramount to its widespread adoption, it is even more important that the training community have faith that algorithms are being trained on validated performance measures. AI can perform many functions, but AI methods will not deliver the performance metrics that they require as inputs.

New training technologies may help the USAF improve training at the margin but *will not necessarily solve the problem either.* The time horizon for delivering new training systems is too long relative to immediate assessment needs. Furthermore, warfare may change faster than the development pipeline can deliver relevant training systems. Putting proficiency measurement first will help ensure (1) that current training approaches are accurately assessed against priority threats and (2) that new training systems will be developed to allow performance assessment (that is, they will include requirements for data capture, data analytic, and scenario generation capabilities, among others).

While the USAF has invested in the development of pilot competencies, *the development of additional competencies will not solve this problem*. The issue with the existing set of competencies is not necessarily that they are wrong, although they might be; rather, it is the lack of performance metrics tied to the competencies and the lack of experimentation and validation of performance metrics.[37] Validation and experimentation could help spur more widespread adoption by demonstrating value through increasing in training effectiveness and/or lowering resource expenditure.

Conclusion

The USAF recognizes the need to prepare its airmen for the complex demands of high-end conflict. The challenges to assessing readiness for these scenarios are considerable and require a coherent strategy across the training enterprise. This Perspective highlights how getting proficiency measurement right is a precondition for achieving this vision and solving persistent challenges across many aspects of the training enterprise. It also highlights the importance of moving from a piecemeal to a holistic approach to managing the training enterprise.

Notes

[1] As we discuss later, *capability assessment* relates to the way that the USAF assesses and reports on the readiness of units to perform tasks, while *proficiency* relates to the skill of individual operators at performing tasks.

[2] Air Combat Command, *Future Training Concept—2020*, November 25, 2019.

[3] Emmi Yonekura, David Schulker, Irina A. Chindea, Ajay K. Kochhar, Andrea M. Abler, Mark Toukan, and Matthew Walsh, *Air Force Readiness Assessment for the High-End Fight: How Training Infrastructure Can Provide Better Information for Decisionmaking*, RAND Corporation, forthcoming.

[4] See U.S. Government Accountability Office, "Air Force Actions to Address Congressionally Mandated Study on Combat Aircrew Proficiency," GAO-20-91, February 2020, for the findings of this assessment.

[5] U.S. Government Accountability Office, 2020, p. 7.

[6] AFI 11-202, Vol. 1, *Aircrew Training*, November 22, 2010, reissued and updated June 10, 2019, p. 24.

[7] For example, AFI 11-2F-16, Vol. 1, defines *proficiency* as "Aircrew have a thorough knowledge of mission area but occasionally may make an error of omission or commission. Aircrew are able to operate in a complex, fluid environment and are able to handle most contingencies and unusual circumstances. Proficient aircrew are prepared for mission tasking on the first sortie in theater" (AFI 11-2F-16, Vol. 1, *F-16—Aircrew Training*, June 17, 2019, Change 2, August 8, 2022, p. 53). while AFI 11-2B-2, Vol. 1, defines it as "Demonstrated ability to successfully accomplish tasked event safely and effectively. For purposes of this volume, proficiency also requires currency in the event, if applicable" (AFI 11-2B-52, Vol. 1, *B-52—Aircrew Training*, September 8, 2011, p. 96).

[8] Department of the Air Force Policy Directive 365-26, *Total Force Development and Management*, April 15, 2022, p. 11.

[9] AFI 11-2F-16, Vol. 2, 2019, p. 46.

[10] Note that objective data, e.g., from simulators or training events, can be used to inform subjective judgment (Jeffrey M. Beaubien, Michael Tolland, and Jared Freeman, "Performance Measurement Applications and Associated Data Requirements for LVC Training," presented at the Interservice/Industry Training, Simulation, and Education 2020 Conference, December 2020).

[11] John A. Ausink, William W. Taylor, James H. Bigelow, and Kevin Brancato, *Investment Strategies for Improving Fifth-Generation Fighter Training*, RAND Corporation, TR-871-AF, 2011.

[12] George M. Alliger, Rebecca Beard, Winston Bennett, Jr., Charles M. Colegrove, and Michael Garrity, *Understanding Mission Essential Competencies as a Work Analysis Method*, Group for Organizational Effectiveness, 2007.

[13] For example, see Naomi C. Brownstein, Thomas A. Louis, Anthony O'Hagan, and Jane Pendergast, "The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making," *American Statistician*, Vol. 73, No. S1, 2019, and Yaniv, 1997.

[14] See Air Force Manual 11-202, Vol. 2, *Aircrew Standardization and Evaluation Program*, August 30, 2021, and AFI 10-201, *Force Readiness Reporting*, December 22, 2020.

[15] Gilroy and Harris, 2020.

[16] Smith and Julia, 2020.

[17] Schreiber, 2013, pp. 294–307.

[18] Mark Toukan, Matthew Walsh, Ajay Kochhar, Emmi Yonekura, and David Schulker, *Air Force Operational Training and Test Infrastructure: Barriers for Full Implementation*, RAND Corporation, RR-A992-1, 2022.

[19] Todd Harrison, "Rethinking Readiness," *Strategic Studies Quarterly*, Vol. 8, No. 3, Fall 2014; Mark Schroeder, Brian Schreiber, and Winston Bennett, "Using Objective Performance Assessments in Applied Settings," in Christopher Best, George Galanis, James Kerry, and Robert Sottilare, eds., *Fundamental Issues in Defense Training and Simulation*, CRC Press, 2017. Some limited experimentation has been done using PETS (Leah J. Rowe, Roger W. Schvaneveldt, and Winston Bennett, Jr., "Measuring Pilot Knowledge in Training: The Pathfinder Network Scaling Technique," Interservice/Industry Training, Simulation, and Education Conference 2007, November 2007).

[20] U.S. Government Accountability Office, *Air Force Training: Further Analysis and Planning Needed to Improve Effectiveness*, GAO 16-864, September 2016.

[21] For approaches that use surveys and interviews as inputs for a model, see Ausink et al., 2011; James H. Bigelow, William W. Taylor, S. Craig Moore, and Brent Thomas, *Models of Operational Training in Fighter Squadrons*, MR-1701-AF, RAND Corporation, 2003; and Sander J. Heis, "Lightning in a Model: Evolving the Training Mix for the F-35," Air University, June 2018.

[22] Kevin A. Gluck, Tiffany S. Jastrzembski, and Michael A. Krusmark, "Prospective Comments on Performance Prediction for Aviation Psychology," Michael A. Vidulich and Pamela S. Tsang, eds., *Improving Aviation Performance Through Applying Engineering Psychology*, CRC Press, 2019.

[23] U.S. Government Accountability Office, 2016.

[24] Susan G. Straus, Matthew W. Lewis, Kathryn Connor, Rick Eden, Matthew E. Boyer, Timothy Marler, Christopher M. Carson, Geoffrey E. Grimm, and Heather Smigowski, *Collective Simulation-Based Training in the U.S. Army: User Interface Fidelity, Costs, and Training Effectiveness*, RAND Corporation, RR-2250-A, 2019.

[25] Straus et al., 2019.

[26] Nick Yates, "OTTI Update: Synthetic Test and Training Capability," briefing slides, HAF/A3TI, undated.

[27] AFI 10-201, 2020.

[28] Yonekura et al., forthcoming.

[29] AFRL's Mission Essential Competencies do capture individual pilot competencies that involve teamwork, although they are defined at the level of individual, not the team.

[30] For a recent model of pilot team performance and a review of existing measurement techniques, see Heikki Mansikka, Kai Virtanen, Donald Harris, and M. Jalava, "Measurement of Team Performance in Air Combat–Have We Been Underperforming?" *Theoretical Issues in Ergonomics Science*, Vol. 22, No. 3, 2021.

[31] Muharrem Mane, Anthony D. Rosello, Paul Emslie, Thomas Edward Goode, Henry Hargrove, and Tucker Reese, *Developing Operationally Relevant Metrics for Measuring and Tracking Readiness in the U.S. Air Force*, RAND Corporation, RR-A315-1, 2020.

[32] U.S. Department of Defense, *Summary of the 2018 National Defense Strategy of the United States of America: Sharpening the American Military's Competitive Edge*, 2018, p. 3.

[33] Charles Q. Brown and David H. Berger, "Redefine Readiness or Lose," webpage, War on the Rocks, March 15, 2021.

[34] For more information on these efforts, see Defense Advanced Research Projects Agency, "AlphaDogfight Trials Foreshadow Future of Human-Machine Symbiosis," press release, August 26, 2020.

[35] Some work has been done on automating performance assessment for certain Navy aviation platforms. See Gilroy and Harris, 2020.

[36] Brian W. Everstine, "CSAF Outlines the Air Force's New Deployment Model," *Air & Space Forces Magazine*, August 4, 2021.

[37] Note that AFRL has begun to identify measures for competencies (and their associated missions, tasks, and skills), as well as candidate measurements that could be taken from aircraft and simulators evaluate performance.

References

AFI—*See* Air Force Instruction.

Air Combat Command, *Future Training Concept—2020*, November 25, 2019.

Air Force Instruction 10-201, *Force Readiness Reporting*, December 22, 2020.

Air Force Instruction 11-2B-52, Vol. 1, *B-52 Aircrew Training*, September 8, 2011.

Air Force Instruction 11-2F-16, Vol. 2, *F-16—Aircrew Evaluation Criteria*, February 8, 2019.

Air Force Instruction 11-202, Vol. 1, *Aircrew Training*, November 22, 2010, reissued and updated June 10, 2019.

Air Force Manual 11-2F-16, Vol. 1, *F-16—Aircrew Training*, June 17, 2019, Change 2, August 8, 2022.

Air Force Manual 11-202, Vol. 2, *Aircrew Standardization and Evaluation Program*, August 30, 2021.

Ausink, John A., William W. Taylor, James H. Bigelow, and Kevin Brancato, *Investment Strategies for Improving Fifth-Generation Fighter Training*, RAND Corporation, TR-871-AF, 2011. As of October 21, 2022: https://www.rand.org/pubs/technical_reports/TR871.html

Alliger, George M., Rebecca Beard, Winston Bennett, Jr., Charles M. Colegrove, and Michael Garrity, *Understanding Mission Essential Competencies as a Work Analysis Method*, Group for Organizational Effectiveness, 2007.

Beaubien, Jeffrey M., Michael Tolland, and Jared Freeman, "Performance Measurement Applications and Associated Data Requirements for LVC Training," presented at the Interservice/Industry Training, Simulation, and Education 2020 Conference, December 2020. As of October 21, 2022:
https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2020&AbID=74237&CID=572#View

Bergenthal, Jeff, William Brobst, Rodney Yerger, and Garrett A. Loeffelman, "Quantifying Future Return on Investment of Live, Virtual, Constructive Training," presented at the Interservice/Industry Training, Simulation, and Education 2020 Conference, December 2020. As of October 21, 2022:
https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2020&AbID=79322&CID=572#View

Bigelow, James H., William W. Taylor, S. Craig Moore, and Brent Thomas, M*odels of Operational Training in Fighter Squadrons*, MR-1701-AF, RAND Corporation, 2003. As of October 21, 2022: https://www.rand.org/pubs/monograph_reports/MR1701.html

Brown, Charles Q., and David H. Berger, "Redefine Readiness or Lose," webpage, War on the Rocks, March 15, 2021. As of November 19, 2021: https://warontherocks.com/2021/03/redefine-readiness-or-lose/

Brownstein, Naomi C., Thomas A. Louis, Anthony O'Hagan, and Jane Pendergast, "The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making," *American Statistician*, Vol. 73, No. S1, 2019.

Defense Advanced Research Projects Agency, "AlphaDogfight Trials Foreshadow Future of Human-Machine Symbiosis," press release, August 26, 2020. As of October 24, 2022: https://www.darpa.mil/news-events/2020-08-26

Department of the Air Force Policy Directive 365-26, Total Force Development and Management, April 15, 2022.

Everstine, Brian W., "CSAF Outlines the Air Force's New Deployment Model," *Air Force Magazine*, August 4, 2021.

Gilroy, Brad, and Dave Harris, "Performance Assessment Using Individual Skills Linked to Mission Outcomes," presented at the Interservice/Industry Training, Simulation, and Education 2020 Conference, December 2020. As of October 21, 2022: https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2020&AbID=76718&CID=572#View

Gluck, Kevin A., Tiffany S. Jastrzembski, and Michael A. Krusmark, "Prospective Comments on Performance Prediction for Aviation Psychology," Michael A. Vidulich and Pamela S. Tsang, eds., *Improving Aviation Performance Through Applying Engineering Psychology*, CRC Press, 2019.

Harrison, Todd, "Rethinking Readiness," *Strategic Studies Quarterly*, Vol. 8, No. 3, Fall 2014.

Heis, Sander J., "Lightning in a Model: Evolving the Training Mix for the F-35," Air University, June 2018.

Mane, Muharrem, Anthony D. Rosello, Paul Emslie, Thomas Edward Goode, Henry Hargrove, and Tucker Reese, *Developing Operationally Relevant Metrics for Measuring and Tracking Readiness in the U.S. Air Force*, RAND Corporation, RR-A315-1, 2020. As of October 21, 2022: https://www.rand.org/pubs/research_reports/RRA315-1.html

Mansikka, Heikki, Kai Virtanen, Donald Harris, and M. Jalava, "Measurement of Team Performance in Air Combat–Have We Been Underperforming?" *Theoretical Issues in Ergonomics Science*, Vol. 22, No. 3, 2021.

Rowe, Leah J., Roger W. Schvaneveldt, and Winston Bennett, Jr., "Measuring Pilot Knowledge in Training: The Pathfinder Network Scaling Technique," Interservice/Industry Training, Simulation, and Education Conference 2007, November 2007.

Schreiber, Brian T., "Transforming Training: A Perspective on the Need for and Payoffs from Common Standards," *Military Psychology*, Vol. 25, No. 3, 2013.

Schroeder, Mark, Brian Schreiber, and Winston Bennett, "Using Objective Performance Assessments in Applied Settings," in Christopher Best, George Galanis, James Kerry, and Robert Sottilare, eds., *Fundamental Issues in Defense Training and Simulation*, CRC Press, 2017.

Smith, Roger, and Danielle Julia, "Machine Learning as an Effective New Tool for Assessing Human Performance During Simulation-based Training," Interservice/Industry Training, Simulation, and Education Conference 2020, December 2020. As of October 24, 2022: https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2020&AbID=75365&CID=572#View

Straus, Susan G., Matthew W. Lewis, Kathryn Connor, Rick Eden, Matthew E. Boyer, Timothy Marler, Christopher M. Carson, Geoffrey E. Grimm, and Heather Smigowski, *Collective Simulation-Based Training in the U.S. Army: User Interface Fidelity, Costs, and Training Effectiveness*, RAND Corporation, RR-2250-A, 2019. As of October 21, 2022: https://www.rand.org/pubs/research_reports/RR2250.html

Toukan, Mark, Matthew Walsh, Ajay Kochhar, Emmi Yonekura, and David Schulker, *Air Force Operational Training and Test Infrastructure: Barriers for Full Implementation*, RAND Corporation, RR-A992-1, 2022. As of October 24, 2022: https://www.rand.org/pubs/research_reports/RRA992-1.html

U.S. Department of Defense, *Summary of the 2018 National Defense Strategy of the United States of America: Sharpening the American Military's Competitive Edge*, 2018.

U.S. Government Accountability Office, *Air Force Training: Further Analysis and Planning Needed to Improve Effectiveness*," GAO 16-864, September 2016.

U.S. Government Accountability Office, *Air Force Actions to Address Congressionally Mandated Study on Combat Aircrew Proficiency*, GAO-20-91, February 2020.

Yaniv, Ilan, "Weighting and Trimming: Heuristics for Aggregating Judgments Under Uncertainty," *Organizational Behavior and Human Decision Processes*, Vol. 69, No. 3, 1997.

Yates, Nick, "OTTI Update: Synthetic Test and Training Capability," briefing slides, HAF/A3TI, undated. As of October 24, 2022: https://www.ntsa.org/-/media/sites/ntsa/events/2021/11a0/slides/haf-a3-part-1.ashx

Yonekura, Emmi, David Schulker, Irina A. Chindea, Ajay K. Kochhar, Andrea M. Abler, Mark Toukan, and Matthew Walsh, *Air Force Readiness Assessment for the High-End Fight: How Training Infrastructure Can Provide Better Information for Decisionmaking,* RAND Corporation, forthcoming.

About the Authors

Mark Toukan is a political scientist at RAND. His research focuses on defense issues, with an emphasis on military training, modeling and simulation, military alliances, and the application of quantitative methods to address policy challenges in these areas. Toukan holds a Ph.D. in political science.

David Schulker is a senior policy researcher at RAND. His research focuses on applying econometric and statistical techniques to help Department of Defense and Department of Homeland Security clients analyze human resource management challenges related to outreach, recruiting, career progression, talent management, workforce development, and retention.

About This Perspective

In this Perspective, we define different dimensions of the U.S. Air Force (USAF) Operational Training and Testing Infrastructure (OTTI) and assess the current state of technologies across these dimensions. The USAF needs to make near-term investment decisions to develop training capabilities so that airmen are ready for high-end conflict in future contingencies. Yet, despite their pressing nature, the USAF lacks needed information to make informed decisions for many training investments. Insufficient research is available to indicate which investments are most appropriate for the most pressing training needs and which should receive priority. This Perspective identifies five primary areas where additional research and information is needed to support investment decisions. We highlight proficiency measurement as key to solving persistent challenges across many aspects of the training enterprise.

This work was conducted within the Workforce, Development, and Health Program of RAND Project AIR FORCE as part of a fiscal year 2021 project "The Use of Operational Training Infrastructure—Live, Virtual, and Constructive Environments in Support of a Squadron Commander's Assessments of Unit Readiness Reporting." A companion report from the same project, *Air Force Readiness Assessment for the High-End Fight: How Training Infrastructure Can Provide Better Information for Decisionmaking*, examines and characterizes shortfalls in the readiness assessment process and then review potential OTTI remedies.

RAND Project AIR FORCE

RAND Project AIR FORCE (PAF), a division of the RAND Corporation, is the Department of the Air Force's (DAF's) federally funded research and development center for studies and analyses, supporting both the United States Air Force and the United States Space Force. PAF provides the DAF with independent analyses of policy alternatives affecting the development, employment, combat readiness, and support of current and future air, space, and cyber forces. Research is conducted in four programs: Strategy and Doctrine; Force Modernization and Employment; Resource Management; and Workforce, Development, and Health. The research reported here was prepared under contract FA7014-16-D-1000.

Additional information about PAF is available on our website: www.rand.org/paf/

Acknowledgments

The authors are grateful to John Ausink for constructive feedback and for the thoughtful reviews of Leah Rowe and RAND colleagues Tim Marler and Anna Jean Wirth. We are also grateful to Barbara Bicksler for her contributions to the structure and design of this Perspective.

RAND CORPORATION

www.rand.org

For more information on this publication, visit www.rand.org/t/PEA992-3.