The 94th MORS Symposium
"Sharpening Our Analytical Edge"

Proposed Framework for Agentic AI Assessment of Creative Thinking for the U.S. Air Force[1] [2]

Yigal Rosen[3], Ph.D.
Co-Founder & Chief Product Officer, Ignis AI

Ilia Rushkin, Ph.D.
Vice President of AI & Data Science, Ignis AI

Invited Paper to be Presented on June 9, 2026 at The 94th Military Operations Research Society (MORS) Symposium "Sharpening Our Analytical Edge" at the USAF Academy, Colorado Springs, CO.

---

[1] The authors would like to express their gratitude to Gene Keselman, Col. (Ret.), MIT Sloan Lecturer and a former USAF Chief Innovation Strategist at the Office of the Vice Chief of Staff (Reserves), as well as the MIT Executive MBA candidate, Dr. Regan Patrick, Lt. Col. (Ret), former Associate Professor, Warrior/Scholar at the USAF Profession of Arms Center of Excellence (PACE), for their invaluable inputs on the earlier draft of this paper.

[2] This research paper is an extended version of the draft research paper and poster presented at the 2026 Artificial Intelligence Workshop: Accelerating Data and Analytics Capabilities for AI Military Operations Research Society (MORS), Carnegie Mellon University.

[3] Correspondence concerning this white paper should be addressed to Dr. Yigal Rosen. Email: yigal@ignisai.ai

Abstract

Agentic AI enables continuous readiness sensing and cognitive readiness development across military training pipelines. This paper proposes a psychometrically validated, agentic AI assessment framework for measuring mission-ready creative thinking competency in the United States Air Force (USAF), with direct implications for Air Education and Training Command (AETC) priorities supporting the 2026 National Defense Strategy (NDS). In contested, AI-enabled environments where adversaries adapt faster than doctrine evolves, mission success depends on Airmen's ability to generate non-obvious options, challenge flawed assumptions, and adapt decisions under pressure faster than the adversary. While USAF competency models identify creative thinking as foundational to readiness, current assessment approaches rely primarily on self-report measures and narrowly scoped simulations that fail to capture real-world performance. Building on the Ignis AI power skills assessment framework, the underlying competencies and skills ontology, and prior research on large-scale assessment of creative thinking, this paper introduces a scenario-based, agentic AI approach that operationalizes creative thinking as observable behavior in mission-relevant contexts. Airmen engage in dynamic, multi-turn interactions requiring problem framing, generation of non-obvious alternatives, and adaptive reasoning under evolving constraints. Responses are evaluated using a human–AI hybrid scoring model grounded in computational psychometrics, incorporating distribution-aware methods to distinguish human creative judgment from AI-generated fluency. The framework aligns with AETC's emphasis on transforming learning, developing mission-ready Airmen, and strengthening decision advantage in AI-enabled environments. By embedding assessment within training pipelines, it enables continuous measurement of cognitive readiness, supports personalized development, and provides scalable analytics for talent management. The paper outlines task design, scoring methodology, and a proposed implementation model for integration into USAF training systems, including pre/post assessment and linkage to performance indicators. This work contributes a scalable readiness infrastructure for developing and sustaining adaptive warfighting capability in the era of AI-enabled warfare.

1. Introduction

The accelerating integration of artificial intelligence (AI) into operational, analytic, and decision-making processes is reshaping both the character of modern warfare and the nature of creativity and innovation required to execute missions successfully. In the United States Air Force (USAF), technological superiority has long been a cornerstone of strategic advantage, yet senior leaders increasingly emphasize that mission success depends as much on human judgment, adaptability, and creativity as it does on advanced systems. As operational environments become more complex, contested, and uncertain, the ability of professionals and leaders to inspire others, organize coordinated action, and generate mission-viable alternatives when standard operating procedures break down.

The USAF competency model explicitly identifies creative thinking as foundational mission-ready capability (Department of the Air Force, 2025). Creative thinking is defined as the capacity to develop new insights in novel situations and to question conventional approaches. Importantly, these definitions position creativity not as abstract traits or positional authority, but as enacted capabilities that integrate cognition, communication, ethics, and social coordination in real operational contexts. Despite its operational importance, cognitive readiness for creative decision-making remains difficult to observe rigorously and at scale. Existing approaches within both military and civilian systems rely heavily on self-report surveys, retrospective evaluations, or narrowly scoped simulations. Such methods are limited in their ability to capture how airmen actually diagnose problems, frame tradeoffs, communicate intent, and adapt their actions as situations evolve. From an operational readiness perspective, these approaches conflate perception with performance and outcomes with underlying capability, reducing their predictive validity in dynamic, high-stakes environments (Rosen, 2015; Rosen et al., 2023). In the USAF high-stakes contexts, limited predictive validity means the assessment of skills fails to predict how warfighters will respond when communications degrade, plans break, and adversaries exploit predictable patterns. These limitations have become more consequential with the rapid diffusion of generative AI systems. Outputs that once served as proxies for human creativity—written analyses, strategic plans, or ideation artifacts—can now be partially or fully generated by AI. While such tools can enhance productivity and baseline output quality, they also introduce substantial ambiguity regarding attribution and capability. Recent research demonstrates that generative AI may increase individual creative output while simultaneously reducing collective diversity, a phenomenon associated with convergence toward high-probability solutions and the erosion of differentiated thinking (Doshi & Hauser, 2024; Kleinberg & Raghavan, 2021; Wu et al., 2025). In readiness sensing contexts, this raises a critical challenge: distinguishing underlying human creative judgment from AI-enabled fluency in ways that are fair, scalable, and aligned with real-world high-stakes performance demands. This paper addresses these challenges by proposing a psychometrically

validated, agentic AI assessment framework for measuring mission-ready creative thinking, anchored in the USAF context but designed to generalize across military contexts and talent development systems facing comparable complexity. Building on Ignis AI's Human Power Skills Ontology, advances in creativity measurement under generative AI, and learning engineering principles, the framework operationalizes creative thinking as observable performance rather than inferred traits. Using agentic AI–driven, scenario-based tasks and distribution-aware, competitive scoring, the approach captures how leaders reason, adapt, and differentiate their thinking under evolving constraints (Baker et al., 2022; Craig et al., 2025). Aligned with the Air Education and Training Command (AETC) Lines of Effort supporting the 2026 National Defense Strategy, this framework addresses a critical gap in the Air Force's ability to develop mission-ready Airmen at scale (Air Education and Training Command, 2026). While AETC is advancing toward data-driven, adaptive training systems, current readiness sensing approaches remain limited in their ability to capture how Airmen think, adapt, and make decisions in complex, uncertain environments. The proposed agentic AI readiness framework enables AETC to operationalize cognitive readiness and decision advantage as continuously observable capabilities. By distinguishing human creative judgment from AI-generated fluency and embedding assessment within learning systems, this approach provides a scalable foundation for continuous readiness analytics, leadership development, and talent management across the force.

Table 1: Alignment to Air Education and Training Command (AETC) Lines of Effort

AETC Line of Effort	How This Framework Directly Supports It
Transform Learning	Scientifically backed scenario-based adaptive assessment embedded in training pipelines
Develop Mission-Ready Airman	Measures real decision behavior, not self-report
Accelerate Change	Enables rapid feedback loops and skill visibility
Strengthen Warfighting Readiness	Identifies gaps in decision-making before mission failure

The contribution of this work is threefold. First, it introduces a practical, scientifically grounded model for assessing creative thinking explicitly aligned with USAF definitions of mission readiness (Department of the Air Force, 2025). Second, it extends emerging theory on creativity under generative AI into applied readiness sensing contexts, demonstrating how distributional and competitive metrics can mitigate AI-driven homogenization and preserve meaningful signals of human capability (Kleinberg & Raghavan, 2021; Raghavan, 2025). Third, it illustrates how learning engineering can translate insights from cognitive science, psychometrics, and AI research into scalable readiness infrastructure systems that support decision-making in high-stakes defense and workforce environments.

2. Creative Thinking in the U.S. Air Force and Modern Organizations

Creative thinking plays a complementary and enabling role in effective leadership under such conditions. The Army Talent Attribute Framework broadly defines creative problem solving as the ability to develop and utilize new or novel and useful methods and strategies to accomplish work or achieve goals in both unexpected, unique or infrequent situations and in evolving and new work environments (U.S. Army Research Institute, 2023). The USAF definition emphasizes developing new insights in novel situations and questioning conventional approaches (Department of the Air Force, 2025), capabilities that are essential when standard operating procedures prove insufficient or when adversaries exploit predictable patterns. Importantly, creative thinking in this context is not synonymous with unconstrained ideation or artistic expression. Rather, it involves disciplined innovation within constraints, reframing problems to surface non-obvious options, and adapting strategies as operational conditions shift.

These competencies are not unique to military contexts. Civilian organizations operating in sectors such as energy, healthcare, finance, and technology face analogous challenges as they navigate regulatory complexity, technological disruption, competitive pressure, and ethical tradeoffs. Leaders in these domains must similarly integrate data, human judgment, and social coordination across complex systems. As in defense, success depends less on technical optimization alone than on the ability to align people, interpret uncertainty, and adapt action in real time.

Despite this alignment across sectors, readiness sensing practices in both military and civilian settings have struggled to keep pace with evolving demands (National Academies of Sciences, Engineering, and Medicine, 2024). Self-report instruments capture dispositions or self-perceptions rather than enacted behavior and are vulnerable to social desirability and reference bias. Retrospective evaluations often conflate outcomes with underlying skill, failing to account for context, chance, or structural constraints. Narrow task-based tests may isolate specific cognitive functions, but rarely capture the integrative

nature of creative thinking as they unfold in practice (Rosen, 2015; Rosen et al, 2020; Rosen et al., 2023).

Scenario-based readiness sensing activities offer a promising alternative by situating individuals in contexts that elicit the behaviors of interest. When designed carefully, scenarios can reveal how leaders frame problems, communicate intent, adapt to new constraints, and organize collective action. However, traditional scenario-based assessments often rely on static prompts and manual scoring, limiting scalability and sensitivity to process-level evidence. From a learning engineering perspective, this constrains the ability to iteratively improve assessment design based on data and to link observed behaviors to development pathways (Baker et al., 2022; Craig et al., 2025).

3. Creativity Measurement in the Age of Generative AI

Creativity has long been recognized as a driver of innovation, adaptability, and strategic advantage, yet its measurement has historically been indirect and contested. Traditional approaches have treated creativity either as a stable trait inferred from personality measures or as a property of isolated outputs evaluated against subjective rubrics. Such approaches were already limited in predictive validity and are increasingly inadequate in AI-mediated environments, where output quality alone is no longer a reliable signal of human capability.

Recent empirical work highlights a paradox introduced by generative AI (e.g., Runco, 2023; Vinchon et al., 2023; Wingström et al., 2024). While AI tools can enhance individual productivity and apparent creativity, they may simultaneously reduce collective diversity by encouraging convergence around high-probability ideas. Doshi and Hauser (2024) demonstrate that generative AI can increase individual creative performance while diminishing population-level variation, a dynamic that undermines long-term innovation. Related work characterizes this phenomenon as algorithmic monoculture, in which widespread reliance on similar models and defaults leads to homogenization of behavior and outcomes (Kleinberg & Raghavan, 2021; Wu et al., 2025).

From a computational perspective, this convergence reflects well-documented properties of generative models. Neural text generation systems tend to collapse toward dominant modes unless explicitly constrained or incentivized to diversify, a pattern observed across language modeling research (Holtzman et al., 2020; Li et al., 2024). Even when prompts request originality, AI-generated outputs often cluster around high-probability regions of the idea space, producing responses that appear novel in isolation but are highly redundant at scale.

In response, emerging theory reconceptualizes creativity as a distributional property rather than a binary attribute or isolated outcome (Acar, 2025; Amabile, & Pratt, 2016). Under this view, creativity is reflected in how ideas are distributed across a conceptual space under shared constraints and incentives (Kleinberg & Raghavan, 2021; Raghavan, 2025). Key dimensions include distinctiveness relative to peers, within-agent variability, sensitivity to contextual changes, and robustness when novelty is explicitly rewarded. Distribution-aware metrics operationalize these dimensions by examining dispersion, divergence, and redundancy across responses rather than relying solely on surface-level quality judgments (Srivastava et al., 2023).

This framing aligns closely with real-world contexts, where creativity is evaluated implicitly through competition for scarce resources, roles, or opportunities. Individuals are rarely assessed on absolute idea quality alone; instead, value is determined by whether ideas offer differentiated advantage relative to alternatives. Crucially, this distributional perspective also provides a principled basis for distinguishing human creative capability from AI-generated or AI-assisted output, which tends to exhibit characteristic clustering patterns even when superficially polished.

The implications for assessment are significant. Quality-only scoring systems—including many rubric-based approaches—are increasingly brittle under generative AI, as they reward fluency and coherence without accounting for convergence or redundancy. Distribution-aware metrics, by contrast, capture properties of ideation and reasoning that remain informative even when surface-level quality is inflated by AI assistance.

This perspective suggests that creative thinking should be evaluated not as a static trait or one-off performance, but as a pattern of behavior across evolving situations. Leaders demonstrate creativity by reframing problems, generating non-obvious options, adapting to new constraints, and sustaining differentiated judgment over time. Capturing these dynamics requires assessment systems that are interactive, comparative, and sensitive to process-level evidence—capabilities that agentic, AI-enabled assessment architectures are uniquely positioned to support.

4. The Ignis AI PowerSkillsAssessment™ Framework

To operationalize creative thinking as mission-ready capabilities, this work builds on the Ignis AI PowerSkillsAssessment™ framework (Rosen, & Rushkin, 2025), a performance-based, AI-enabled system designed to provide force-wide readiness visibility into critical human capabilities at scale. The framework is grounded in a Human Power Skills Ontology that defines creativity not as latent traits inferred from self-report, but as patterns of observable behavior enacted across realistic, high-stakes contexts.

At its core, the framework reflects a learning engineering approach to assessment design, integrating theory, task construction, scoring, and iteration within a coherent system (Rosen, & Rushkin, 2026a; Rosen, & Rushkin, 2026b). Rather than treating assessment as a static instrument, the framework is designed as an adaptive measurement system in which design decisions are continuously informed by empirical evidence and aligned with operational use cases, including selection, development, and readiness analytics.

4.1 Human Power Skills Ontology and Creative Thinking Construct

The Ignis AI Human Power Skills Ontology specifies creative thinking as a multi-dimensional construct composed of interrelated cognitive, social, and ethical capacities. In alignment with the USAF competency model (Department of the Air Force, 2025), creative thinking is operationalized as the capacity to develop new insights, question conventional approaches, and adapt solutions within constraints.

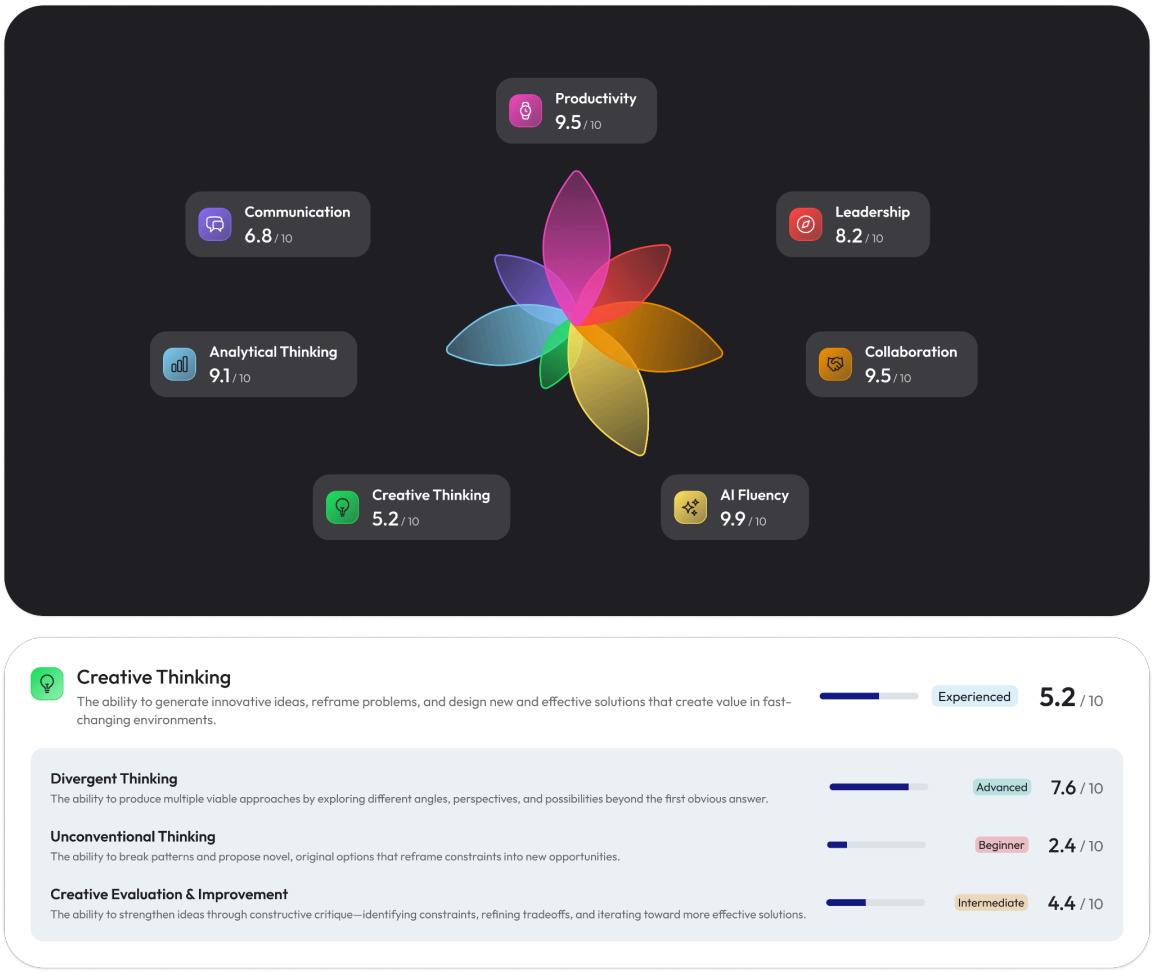


Figure 1. Ignis AI PowerSkillsPrint™: Human Power Skills Talent Flower

These capacities are intentionally defined at a level of abstraction that supports both defense and civilian workforce applications, while remaining anchored in USAF doctrine and operational realities.

Figure 1 illustrates Ignis AI PowerSkillsPrint™, a structured visualization designed to make human power skills visible, interpretable, and actionable. The central "talent flower" represents an individual's multi-dimensional performance profile across seven human power skills assessed by the Ignis AI PowerSkillsAssessment™: Creative Thinking, Leadership, Communication, Collaboration, Analytical Thinking, Productivity, and AI Fluency. Each petal corresponds to a distinct skill domain, with its relative size and numerical proficiency estimate reflecting performance inferred from scenario-based, agentic AI interactions rather than self-report. Proficiency scores are expressed on a calibrated scale and accompanied by uncertainty-aware performance bands, enabling both developmental insight and responsible interpretation for high-stakes contexts. Surrounding the talent flower, the PowerSkillsPrint™ provides deeper diagnostic views for selected competencies, including Creative Thinking, which are central to USAF mission readiness. These expanded views decompose each competency into theoretically grounded sub-skills aligned with operational definitions and learning science. For Creative Thinking, sub-skills include Divergent Thinking (the ability to generate multiple viable approaches), Unconventional Thinking (the capacity to break patterns and propose original options), and Creative Evaluation and Improvement (the ability to refine ideas through critique, tradeoff analysis, and iteration). These dimensions capture creativity as disciplined innovation under constraints, consistent with USAF requirements for developing new insights and questioning conventional approaches in uncertain environments.

Importantly, the PowerSkillsPrint™ integrates performance-level descriptors (e.g., Beginner, Intermediate, Experienced, Advanced) to translate quantitative estimates into interpretable developmental signals. This design supports formative feedback for individual development while maintaining the rigor required for selection and readiness analytics. By grounding all scores in observable behavior elicited through mission-relevant scenarios, the Ignis AI PowerSkillsPrint™ demonstrates how AI-enabled assessment can move beyond abstract potential toward a concrete, evidence-based representation of mission-ready creative thinking.

4.2 Scenario-Based, Agentic AI Task Design

Readiness sensing within the Ignis AI framework is conducted through scenario-based tasks that simulate mission-relevant decision environments. Scenarios are designed to elicit the behaviors specified in the ontology by placing leaders in situations characterized by ambiguity, competing objectives, resource constraints, and ethical tension. Rather than

selecting from predefined options, participants engage in open-ended interaction with agentic AI systems that dynamically respond to their inputs.

This agentic design enables the capture of process-level evidence, including how leaders frame problems, generate options, revise assumptions, and communicate rationale over time. Prior research demonstrates that such human–agent interaction can validly elicit complex skills such as collaboration, communication, and creative problem solving in ways that traditional static tasks cannot (Rosen, 2015; Rosen et al., 2023).

Importantly, scenarios are structured to evolve based on participant responses, introducing new information, constraints, or stakeholder perspectives. This dynamic structure allows assessment of adaptability and creative thinking as situated behavior rather than isolated output, consistent with USAF definition of creativity as enacted capabilities (Department of the Air Force, 2025).

Below are two excerpts (beginnings of scenario conversations) from an Ignis-generated activity focused on evaluating the Creative Thinking skill. The text in grey represents technical background information essential for understanding the key considerations of the assessment design and development process described in this paper. The generation engine was provided with the Air Force handbook (Department of the Air Force, 2025) as a reference material. Upper and lower bounds were imposed on the aggregate complexity score, which is the mean of the additional scores (see examples of those scores below).

The first excerpt is designed as an illustrative example for a baseline assessment at a manager of middle or upper seniority level, whose functional area is “program management”.

> Hi, I’m Colonel Samantha Lewis, Program Director at Air Combat Command. We’re facing an urgent challenge: our fighter squadron maintenance scheduling system is outdated, causing delays, rising costs, and mission readiness gaps. We must modernize across multiple bases, with tight budgets, strict cybersecurity policies, and varied local procedures. Over the next few prompts, I’d like you to help me explore a range of ideas, break conventional patterns, and refine the best concepts so we can deliver an innovative, practical scheduling solution across the command.
>
> First, I’d like you to brainstorm several different approaches we could take to modernize the scheduling system. What options come to mind?


```
"aggregate_complexity_score": 0.5142857142857142,
"additional_scores": {
  "linguistic_complexity": 0.6,
  "conceptual_abstraction": 0.8,
  "number_of_reasoning_steps": 0.4,
  "dependency_among_ideas": 0.2,
```

```
  "degree_of_inferential_reasoning": 0.6,
  "likelihood_of_common_misconceptions": 0.4,
  "working_memory_load": 0.6
}
```


Response examples (for internal use only, *not shown to the participant*):

- (score 0) We could purchase a commercial scheduling software package and install it to replace the legacy tool.
- (score 1) Maybe we could upgrade to a web-based scheduling tool and also standardize our spreadsheet templates across bases.
- (score 2) We could adopt a cloud-based scheduling platform, integrate with flight-line asset tracking, use mobile notifications for crews, or employ simple AI rules to optimize slot assignments.
- (score 3) We could build a modular cloud-native system, leverage machine learning to predict maintenance demand, create a mobile app for on-the-go updates, establish a blockchain log for audit trails, and introduce peer-to-peer scheduling to share capacity across nearby bases.

Skills:

- **Divergent Thinking**. (Primary tag). Rationale: This prompt asks you to generate multiple distinct approaches to a complex problem, directly testing divergent thinking—the ability to explore different angles beyond the first obvious solution.
- **Unconventional Thinking**. Rationale: This prompt explicitly asks to brainstorm multiple novel approaches to modernizing an outdated scheduling system, encouraging them to break conventional patterns across constraints such as budgets, cybersecurity policies, and varied local procedures. It directly tests the ability to generate original, unconventional solutions that reframe existing constraints into opportunities, aligning tightly with the Unconventional Thinking skill definition.

.

Next, consider how we might involve cross-base collaboration in the solution. What different angles or possibilities do you see?


```
"aggregate_complexity_score": 0.5714285714285714,
"additional_scores": {
  "linguistic_complexity": 0.6,
  "conceptual_abstraction": 0.8,
  "number_of_reasoning_steps": 0.8,
  "dependency_among_ideas": 0.2,
  "degree_of_inferential_reasoning": 0.6,
  "likelihood_of_common_misconceptions": 0.4,
  "working_memory_load": 0.6
```

}

Response examples (for internal use only, *not shown to the participant*):

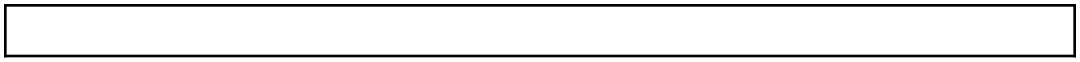
- (score 0) We could ask each base to submit their schedules to my office.
- (score 1) We might set up a shared spreadsheet and hold a monthly conference call for each base to report their schedules.
- (score 2) We could create a centralized cloud dashboard, assign rotating cross-base liaisons, implement a joint scheduling portal, or integrate with the existing Joint Base Scheduling Integration tool.
- (score 3) We could develop an interactive portal with geofencing alerts, appoint scheduling champions at each base, host hackathons to generate local solutions, or establish a federated network that balances load dynamically across nearby installations.

Skills:

- **Divergent Thinking**. (Primary tag). Rationale: This question requires you to generate multiple approaches that leverage collaboration across different units, assessing your comfort with open-ended brainstorming and exploring varied perspectives.
- **Unconventional Thinking**. Rationale: The prompt explicitly asks to break conventional patterns and propose novel, original options for cross-base collaboration under tight constraints. By seeking multiple angles and creative possibilities for a logistics challenge, it tests the ability to reframe limits into new opportunities, directly aligning with the Unconventional Thinking skill.

The second excerpt is designed an an illustrative example for a baseline assessment for cadets in the US Air Force Academy:

> Hi, I'm Captain Sarah Miller, Training Officer at the United States Air Force Academy. We're halfway through a winter field training exercise when a recon drone carrying sensitive intel has gone down in a blizzard. Communications with the forward element have failed, our satellite uplink is offline, and standard recovery procedures can't be executed. You have limited information on enemy movements, severe weather closing in, and only what you can improvise. I need your help generating multiple mission-viable alternatives under these ambiguous conditions.
>
> Let's start broad. How would you generate a range of distinct approaches to locate and secure the downed drone despite the communications breakdown and white-out conditions? List three very different options.



"aggregate_complexity_score": 0.6,

```
"additional_scores": {
  "linguistic_complexity": 0.6,
  "conceptual_abstraction": 0.8,
  "number_of_reasoning_steps": 0.6,
  "dependency_among_ideas": 0.2,
  "degree_of_inferential_reasoning": 0.8,
  "likelihood_of_common_misconceptions": 0.6,
  "working_memory_load": 0.6
}
```


Response examples (for internal use only, *not shown to the participant*):

-
- (score 0) I'd follow our standard recovery SOP: assemble a ground team, navigate to the last known GPS fix, and secure the drone.
- (score 1) We could send a small patrol to grid-search where it went down, or attempt to reestablish comms then dispatch a second team. Otherwise, we might wait for clearer weather before moving in.
- (score 2) One team could use compass and map to perform a grid search; another could climb a ridge for visual spotting and signal with flares; or we could launch a handheld UAV for thermal imaging to pinpoint its location.
- (score 3) I'd split into three distinct efforts: a small ground squad using snowshoes to methodically grid the area; an overwatch team sending up flares and using handheld IR scanners; and a liaison effort to rig a portable satellite uplink from a nearby training bunker for remote drone tracking.
-

Skills:

- **Divergent Thinking**. (Primary tag). Rationale: This item asks the cadet to move beyond the single obvious solution by generating multiple routes to achieve the mission under ambiguous conditions, directly assessing divergent thinking—the skill of exploring varied perspectives and possibilities rather than a linear, single-answer approach.
- **Analysis & Evaluation**. Rationale: This task requires analyzing a highly complex, ambiguous scenario—downed drone in a blizzard with lost communications—and logically evaluating multiple constraints (weather, enemy movements, resources). To propose three distinct mission-viable alternatives, cadets must assess evidence quality, ask critical questions about terrain, weather patterns, and enemy positions, and evaluate potential options to make sound decisions under pressure, directly testing Analysis & Evaluation.
- **Unconventional Thinking**. Rationale: By asking for three very different approaches under severe constraints, the assessment explicitly targets Unconventional Thinking. Cadets must break conventional recovery procedures, reframe the white-out conditions and communications breakdown as

opportunities, and generate novel, original alternatives (e.g., sensor-kite deployment, diversionary tactics, improvised relay networks), demonstrating their ability to think outside standard patterns and innovate solutions in a high-stakes environment.

Now imagine enemy forces might be converging on the site. What additional alternative courses of action would you propose to maximize success and minimize risk? Describe three approaches from different angles.


```
"aggregate_complexity_score": 0.5428571428571429,
"additional_scores": {
  "linguistic_complexity": 0.6,
  "conceptual_abstraction": 0.8,
  "number_of_reasoning_steps": 0.6,
  "dependency_among_ideas": 0.2,
  "degree_of_inferential_reasoning": 0.6,
  "likelihood_of_common_misconceptions": 0.4,
  "working_memory_load": 0.6
}
```


Response examples (for internal use only, *not shown to the participant*):

- (score 0) I'd keep the team small and move fast to avoid detection, then extract immediately.
- (score 1) We could send one squad to create a diversion while another retrieves the drone, or drop smoke to hide our approach, and maybe call for aerial recon if comms come back.
- (score 2) One option is a diversionary flanking patrol to draw eyes away; another is using a snowmobile convoy under camouflage nets; a third is planting motion sensors around the crash site to warn us of enemy movement.
- (score 3) I'd propose three distinct approaches: a nighttime silent infiltration team using crampons and NVGs; a decoy squad deploying infrared flares at a false crash site; and an aerial retrieval using an unmanned tethered glider launched from the academy roof to grab and airlift the drone.

Skills:

- **Divergent Thinking**. (Primary tag). Rationale: By asking the cadet to propose multiple risk-mitigating options under threat, this item measures divergent thinking—encouraging the cadet to explore several distinct strategies rather than relying on a single, conventional plan.
- **Analysis & Evaluation**. Rationale: This scenario requires the cadet to logically assess limited intelligence, weather conditions, and enemy activity, then evaluate multiple recovery options under high ambiguity. Generating

mission-viable alternatives involves analyzing risks, weighing trade-offs, and making sound decisions in a complex, uncertain environment, directly aligning with Analysis & Evaluation.

- **Synthesis**. Rationale: Proposing three distinct approaches demands integrating diverse inputs—weather, enemy movements, communications failures, and improvised resources—into coherent strategies. The cadet must connect these disparate factors to form actionable courses of action, demonstrating the Synthesis skill of unifying varied information into clear, executable plans.
- **Extrapolation**. Rationale: Anticipating enemy convergence, weather shifts, and communication breakdowns compels the cadet to forecast implications and plan proactively under uncertainty. Generating alternatives that preemptively address evolving threats and environmental trends directly tests Extrapolation, as the cadet must use patterns to predict next steps and adapt mission planning accordingly.
- **Unconventional Thinking**. Rationale: The instruction to propose three approaches from different angles under extreme constraints encourages breaking conventional procedures and devising novel tactics. Cadets must reframe standard recovery methods into innovative solutions, exemplifying Unconventional Thinking by generating original courses of action that challenge routine military doctrine.

4.2.1 Baseline Cognitive Readiness Assessment for Academy Cadets

Although illustrative, the cadet scenario highlights a broader strategic application of the framework within accession and officer development pipelines. Institutions such as the U.S. Air Force Academy, West Point, and the U.S. Naval Academy increasingly face the challenge of preparing future officers for operational environments characterized by uncertainty, distributed decision-making, and AI-enabled conflict. Traditional academic indicators and leadership evaluations provide only limited visibility into how cadets generate options, adapt under pressure, and exercise judgment in ambiguous situations.

The proposed framework enables the establishment of baseline cognitive readiness profiles early in the development pipeline. By administering scenario-based assessments at key transition points—such as accession, summer field training, commissioning preparation, and pre-command development—the services can observe how creative thinking, adaptability, and decision-making evolve longitudinally across a cadet's formation experience.

This longitudinal perspective is particularly important because mission-ready creative thinking is developmental rather than static. The framework supports repeated measurement using varied but psychometrically linked scenarios, enabling growth modeling over time while reducing memorization and gaming effects. Such capability creates opportunities to identify both exceptionally adaptive performers and individuals who may require targeted developmental intervention long before operational assignment.

Beyond individual development, aggregated cohort analytics may also provide strategic insight into institutional effectiveness. Military academies and training commands could evaluate whether specific curricula, experiential learning activities, leadership programs, or operational simulations are measurably improving the cognitive and creative capabilities associated with mission readiness. In this sense, assessment becomes not only a learner-level capability, but an institutional learning system for continuous force development.

4.3 Distribution-Aware and Competitive Scoring

Drawing on emerging theory in generative AI and creativity measurement, creativity is treated as a distributional property that emerges under shared constraints and competitive conditions (Kleinberg & Raghavan, 2021; Raghavan, 2025). These conditions are informed by computational work on diversity and mode collapse in language models (Holtzman et al., 2020; Li et al., 2024; Wu et al., 2025) and by distributional measures of generated text diversity (Srivastava et al., 2023). By focusing on patterns of ideation rather than surface-level quality alone, the framework distinguishes human creative judgment from AI-generated or AI-assisted fluency, which tends to cluster around high-probability solutions even when prompts request originality.

4.4 Human–AI Hybrid Scoring and Fairness

Scoring within the Ignis AI framework is implemented through a human–AI hybrid approach. AI models are used to extract features, model distributions, and generate preliminary scores, while human expert oversight is retained for rubric validation, bias monitoring, and adjudication of edge cases. This design supports reliability and scalability while maintaining transparency and alignment with ethical and fairness considerations.

From a learning engineering perspective, this hybrid architecture enables iterative refinement of scoring models based on empirical evidence, supporting continuous improvement and alignment with operational needs (Baker et al., 2022; Craig et al., 2025). It also mitigates risks associated with over-automation, particularly in high-stakes defense and talent decisions where accountability and explainability are essential.

Together, these design elements establish a measurement system capable of capturing creative thinking as dynamic, context-sensitive performance. The next section outlines the proposed research and pilot design for evaluating this framework within USAF training pipelines, including validation strategies and linkage to performance indicators.

5. Task Design, Agentic AI, and Cognitive Readiness Infrastructure

Assessing mission-ready creative thinking competencies requires tasks that reflect how individuals actually operate in real environments rather than simplified testing conditions. Leaders in the USAF and comparable civilian organizations rarely encounter well-structured problems with clear objectives and complete information. Instead, they operate under uncertainty, competing priorities, ethical constraints, and dynamic adversary or market behavior (Department of the Air Force, 2025). To capture these realities, the Ignis AI PowerSkillsAssessment™ relies on scenario-based, multi-turn tasks delivered through agentic AI systems that elicit creative thinking as enacted performance. From a learning engineering perspective, this approach treats assessment as a designed system in which task structure, interaction dynamics, and scoring models are intentionally aligned with the constructs of interest (Baker et al., 2022; Craig et al., 2025). The goal is not to approximate operational complexity perfectly, but to elicit the cognitive, social, and ethical behaviors that underlie creativity in mission-relevant contexts.

5.1 Scenario-Based Task Design

Each assessment scenario is designed to reflect authentic challenges, such as allocating limited resources, responding to unexpected failures, balancing mission objectives with safety and ethics, or communicating intent across hierarchical and functional boundaries. Scenarios unfold over multiple turns, with new information, constraints, or stakeholder perspectives introduced as the participant responds. This structure enables the assessment to capture not only what decisions are made, but how leaders reason, adapt, and communicate as situations evolve. From an assessment science perspective, scenarios are explicitly aligned with a construct model grounded in the Human Power Skills Ontology and USAF competency definitions (Department of the Air Force, 2025). This alignment ensures that observed performance can be interpreted meaningfully rather than treated as an opaque outcome, consistent with prior work on performance-based assessment of complex skills (Rosen, 2015; Rosen et al., 2023).

5.2 Agentic AI as Assessment Orchestrator

A defining feature of the framework is the use of agentic AI not merely as a conversational interface, but as an assessment orchestration layer that continuously

probes, adapts, and models mission-relevant cognitive performance. Unlike static assessments that capture isolated outputs at a single point in time, agentic AI systems dynamically shape the interaction based on participant decisions, evolving constraints, and inferred proficiency signals. In practice, this enables the assessment environment to more closely resemble operational command contexts in which information is incomplete, priorities shift rapidly, and adversaries actively adapt.

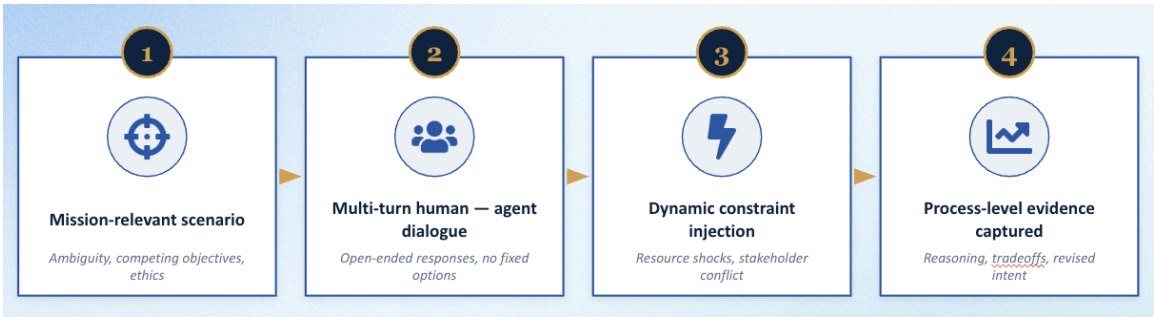


Figure 2. Agentic AI as Assessment Orchestrator

From a learning and readiness perspective, the orchestration capability is particularly significant because it transforms assessment from an episodic evaluation event into a continuous readiness sensing capability embedded within the training pipeline. As Airmen interact with scenarios, the system can identify emerging strengths and weaknesses, escalate or reduce complexity, inject new operational constraints, and generate individualized evidence profiles tied to mission-relevant competencies. This creates a persistent developmental feedback loop in which assessment, coaching, and readiness analytics become tightly integrated rather than institutionally separated. The orchestration model mirrors the realities of AI-enabled warfare itself. Operational leaders increasingly face environments characterized by degraded communications, conflicting intelligence, autonomous systems, and adversaries capable of rapid adaptation. By introducing uncertainty, ambiguity, and evolving tradeoffs into the interaction, agentic AI assessment captures not only whether participants arrive at viable solutions, but how they revise assumptions, communicate intent, and sustain differentiated judgment as conditions change. These process-level signals are often more predictive of mission resilience than the final answer alone (Rosen, 2015). Importantly, agentic AI also supports controlled variation across participants. While all participants engage with scenarios drawn from a common template, specific details may differ to reduce rote responding and limit the effectiveness of AI-only generation. This design choice strengthens fairness and mitigates gaming in AI-mediated assessment contexts, where static prompts are increasingly vulnerable to automation.

5.3 Human–AI Hybrid Scoring and Bayesian Proficiency Estimation

Scoring follows a human–AI hybrid model designed to balance scalability, transparency, and accountability. AI models generate preliminary scores, feature representations, and uncertainty estimates based on trained classifiers and probabilistic models. Human experts oversee rubric calibration, audit scoring behavior, and adjudicate ambiguous cases. This structure aligns with responsible AI principles for high-stakes readiness and talent decisions, and mitigates risks associated with fully automated judgment. Proficiency estimation is conducted using Bayesian modeling, integrating evidence across multiple tasks and interactions. Rather than producing point estimates alone, the system generates posterior distributions that explicitly represent uncertainty. As participants complete additional tasks, uncertainty bands narrow, increasing confidence in proficiency estimates. This approach supports both summative decisions and formative feedback while avoiding overinterpretation of sparse or noisy data, consistent with best practices in performance-based assessment and learning engineering (Rosen et al, 2018; Baker et al., 2022; Craig et al., 2025). Below we present an evaluation of AI model scoring assessments on a human-labeled dataset, for a subset of skills: this is a direct comparison of the scores produced by the AI model and those produced by a human expert ("labels"). The scores are normalized to lie on the 0-to-1 scale. We measure the mean absolute error of the score, and the Pearson correlation between the AI score and the label score.

Table 2: A direct comparison of the scores produced by the AI model and those produced by a human expert ("labels"), for a subset of skills.

Power Skill	Mean Absolute Error (MAE)	Pearson Correlation with Label	Label standard deviation	AI score standard deviation	Quadratic Weighted Kappa (QWK)	Accuracy	n
AI Fluency	0.082	0.737	0.262	0.231	0.731	0.775	285
Analytical Thinking	0.068	0.760	0.221	0.210	0.758	0.797	286
Collaboration	0.077	0.730	0.239	0.217	0.726	0.782	321
Communication	0.067	0.850	0.273	0.271	0.849	0.800	210
Leadership	0.103	0.718	0.257	0.232	0.712	0.692	120
Overall	**0.077**	**0.774**	**0.256**	**0.238**	**0.772**	**0.778**	**1229**

Both the labels and the AI scores are scaled to be on the 0-to-1 scale. Column "Accuracy" is made possible by the fact that in this implementation the model spectrum (all the possible produced score values) was the same as in the labels (0, 1/3, 2/3, 1), so it could be viewed as a classification model as well as a regression model.

We can also compare the distributions of assessment responses by score, using the human-label scores (“label”) and the scores from the AI model (“AI”). These histograms show that the distribution is not heavily skewed, and that difference between the label distribution and the AI score distribution is small.

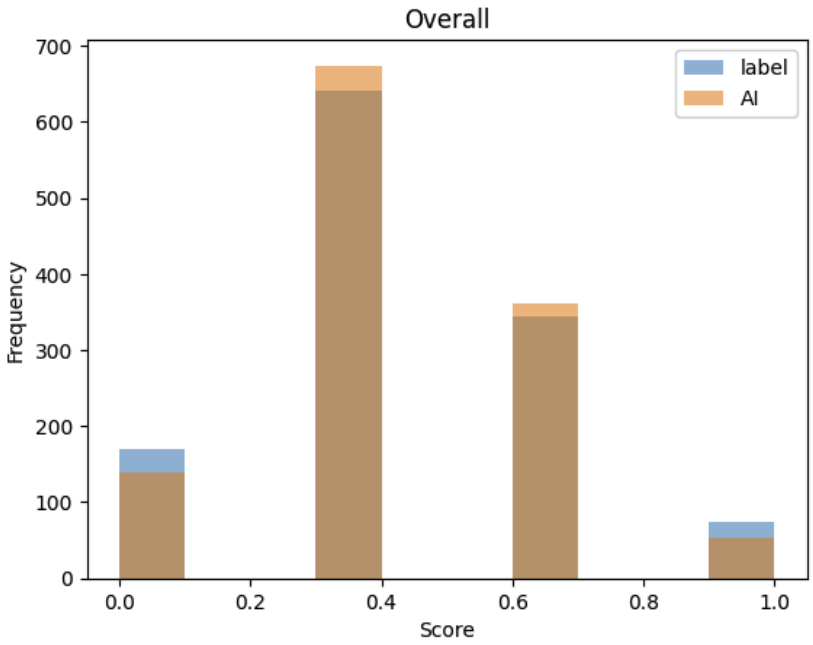


Figure 2. The distributions of assessment responses by score, using the human-label scores (“label”) and the scores from the AI model (“AI”).

Assessment of the skill *Creative Thinking* can follow the same general framework as other skills; however, key aspects of creativity require specialized measurement approaches. One promising approach measuring creativity, which we are currently developing (Rosen & Rushkin, 2026a), relies on modeling creativity as novelty (i.e., mission-viable alternatives) and entropy in synthesis of ideas. Given a set of premise statements (abstract ideas, disparate facts, concepts from different domains, etc.), an inference statement is produced by the test subject in response.

This formulation builds on established theories of creativity, including divergent thinking as the expansion of the idea space (Runco & Acar, 2012) and associative recombination as the basis of creative insight (Mednick, 1962). It also aligns with research on ambidextrous leadership, which emphasizes the interplay between exploratory and exploitative processes in innovation (Rosing et al., 2011).

We implement this framework using a geometric representation in embedding space, where premises and responses, as well as their subelements, are represented as high-dimensional vectors. Considering the projection of a response (or a response

subelement) onto a cone spanned by the premises (or their subelements), novelty is quantified by the norm of the complement to the projection, while transformation is quantified using entropy-based measures that capture the breadth and balance of contributions across premises. The two measures—novelty and entropy—are then combined into a single creativity score by means of a mathematical model that allows tunable meta-parameters. Importantly, this approach does not rely on opaque generative processes or subjective scoring. Instead, it provides a transparent and consistent numerical measure grounded in first principles. While it does not claim to capture all aspects of creativity, it offers a valid and operationalizable definition that can be applied across diverse contexts. We evaluate the proposed framework using a synthetic AI-generated dataset of activities (sets of premises) and responses of varying levels of creativity. This provides us with a labeled set of activities and responses. Generation is done with a general-purpose generative model with straightforward prompts, so that it serves as an approximation to human judgments of what is more or less creative. Thus, while we compare our model scores to the labels, the labels cannot be regarded as absolute truth, only an approximation to human-perceived creativity. Hence we are evaluating not only literal closeness (mean absolute error), but also correlation: Pearson rho and Kendall tau (a measure of ordinal agreement).

Figure 3. Preliminary results demonstrate that creativity can be measured reliably across assessment activities.

Mean absolute errors of 40 activities are typically low (with both the model outputs and the labels being on 0-to-1 scale): 0.20 average across activities. Kendall and Pearson correlation coefficients are typically high (0.61 and 0.76 averages across activities).

Score distributions are provided to illustrate the relationship: on the x-axis are the labels (converted to integers 0-4). The box-and-whiskers plots show the sets of model outputs for those responses. We see that they go steadily up with X and cover a large portion of the theoretically defined 0-1 range on the y-axis.

Results indicate low mean absolute error and substantial correlation between model scores and intuitive labels, suggesting that the framework aligns with common notions of perceived creativity. Moreover, the variability of metrics across activities provides us with further insights and serves as the beginning of an iterative convergent process: analyzing differences between activities with better and worse agreement metrics (like #33 vs. #13), we will iterate on the activity-generation process in order to further stabilize it and reduce the scatter. Tests on human-labeled data will also be done.

5.4 The 2026 Studies: Psychometric properties of generated assessment and scoring

Psychometric properties of the assessment are examined to validate both the AI scoring system and the assessment content, creation of which is becoming more and more AI-assisted. We pay particular attention to fairness and biases. Choosing a respondent grouping variable (such as ethnicity), and selecting an item set developed for the same Power Skill, we set up our Differential Item Functioning (DIF) analysis by modeling the score of a respondent on an item as *score ~ rest_score + group + rest_score:group*. We use a fractional-binomial generalized linear model, separate for each item in the item set. The rest score is defined as the sum of the user's scores on all items except the one being modeled. The fit coefficient for the term group represents *uniform DIF* (a potential bias of the item for or against an entire group, regardless of their overall performance in the item set) , and the coefficient for the interaction term represents the non-uniform DIF (a potential bias of the item that depends on the respondent performance, but also on their group). An item is marked if any of these coefficients is statistically significant, based on a p-value threshold, which we set to 0.1.[4] The model is further used to predict the score using the group-averages of rest scores, so that the differences between predictions represent the gaps among groups. An item is marked if the maximum group gap exceeds a threshold. An item that received one of the marks is flagged for human review but not removal: either statistical evidence is weak but implied consequences are large, or vice versa. An item that received both marks is flagged for removal. The item-level results are then aggregated to the level of the item set.

[4] We would like to remind here that, perhaps counter-intuitively, a higher p-value means a more stringent DIF test. It means that the test will flag even a DIF with weaker evidence.

In a recent study conducted in March-April 2026 with a balanced diverse group of respondents recruited through the Prolific research platform (n=613) who took the full battery of assessment activities that we generated using our hybrid AI+Human SME (Subject Matter Expert) system (54 items in total). The item sets in the analysis were defined by the item's primary association with one of our 7 Power Skills. The DIF analysis was based on **sex/gender, ethnicity (Asian/Black/White), age group (18-40 yo vs. 41-69 yo), employment status (binary) or student status (binary)**. Overall, **no significant difference was found among any groups.** In one item only (associated with the skill "Analytical Thinking"), weak evidence of non-uniform DIF was found with respect to age group: the predicted group average score for the 41-69 age group was slightly higher than for the 18-40 age group (0.514 vs 0.497, a 0.017 difference, with p=0.068). This item was removed.

Furthermore, a 2PL-IRT analysis is applied to estimate the item difficulty and discrimination power. In the same study, no obvious outliers were found, and the item difficulty was found to cover a wide range (0.2-0.6 on the 0-1 scale) and the discrimination was consistently 0.7-0.8 on the 0-1 scale (see Figure 4) .

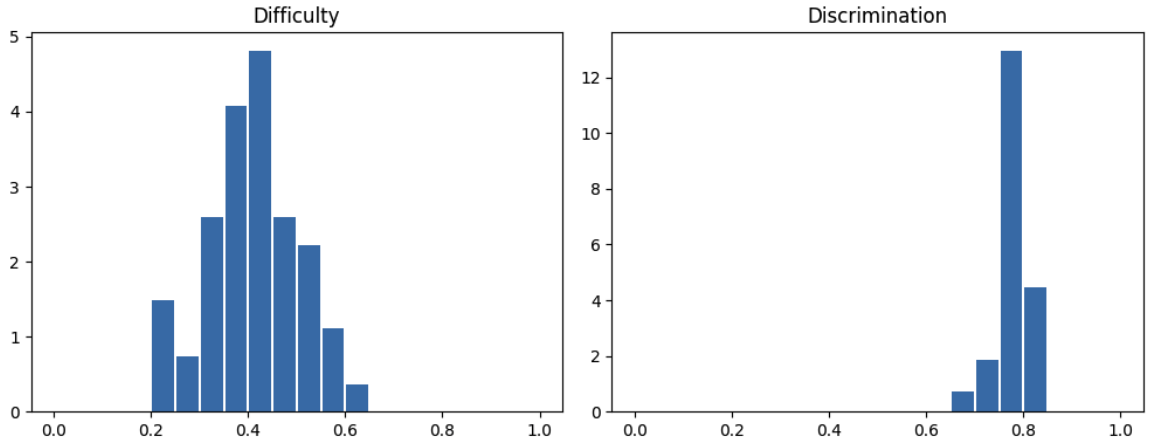


Figure 4. Difficulty and discrimination parameters of 54 AI-generated, SME reviewed items in the study.

A separate model was fitted for each Power Skill. For interpretability, both parameters are transformed after the model to [0, 1] range. For "difficulty", the transformation is the standard normal CDF (same as the population distribution of ability assumed by the IRT model), for "discrimination" it is x/(1+x). The interpretation of "difficulty" values are: 0.15-0.4 is moderately low, 0.4-0.6 is normal, 0.6-0.85 is moderately high. The interpretation of "discrimination" values are: 0.3-0.5 is moderately low, 0.5-0.6 is normal, 0.6-0.8 is moderately high.

In a follow-up study, also conducted in April 2026, the same methodology was used with the content developed to measure Power Skills in the Leadership Edition (40 items across 3 Power Skills, n=431). **No significant difference was found among any groups.** The

IRT parameter distribution was similar to the one found in Study III. The difficulty ranged between 0.2 and 0.8, and the discrimination between 0.7 and 0.85.

Another study was conducted in April 2026 with the goal of evaluating the agreement between human scorers and our AI scoring model. The responses obtained from Study III were used. The study was conducted in a multi-round setup. 20 scorers were initially recruited through the Prolific research platform, screened for management experience, reasoning skills and AI annotation experience. After the first calibration round, 10 were selected based on scoring agreement metrics and results inspection. After the second calibration round, 6 were selected, who then participated in 3 more rounds. In total, 400 responses, randomly selected from the data, were scored, each by the entire group of scorers. The consensus score was formed as the mean score. The results were:[5]

- Inter-rater reliability among human scorers: mean quadratic weighted kappa (QWK) of 0.708 and Krippendorff's alpha of 0.714.
- Consistency between AI scoring and individual human evaluations: mean QWK of 0.723.
- Agreement between AI scoring and the consensus human score: QWK of 0.788.

Substantial inter-rater reliability indicates that the assessment responses are not strongly open to interpretation. Substantial agreement of AI with the human raters indicates that the model evaluates responses in largely the same way as human scorers. We note that the QWK between the mode and a human scorer is on average higher than among different human scorers.

6. Implications for Mission Readiness and Talent Development

The cognitive readiness framework described in this paper has direct implications for how mission-ready performance is identified, developed, and sustained within the U.S. Air Force. More broadly, it offers a generalizable model for organizations operating in complex, high-stakes environments where performance effectiveness depends on creative judgment, ethical reasoning, and coordinated action under uncertainty.

Within the USAF, mission readiness depends not only on technical proficiency and procedural compliance, but on leaders' ability to adapt when plans encounter friction, ambiguity, or adversarial disruption. By operationalizing creative thinking as observable performance aligned with Air Force competency definitions (Department of the Air

[5] For calculating metrics that require the scores to have a certain ordered set of values, rather than any value within a range, the consensus scores of human scorers, as well as the AI model scores, were rounded to the nearest such value.

Force, 2025), the proposed framework enables a more precise understanding of readiness at the individual and unit levels.

The scenario-based, agentic design supports early identification of strengths and gaps in capability before they manifest in operational settings. Because proficiency estimates are probabilistic and updated as additional evidence is collected, the framework supports longitudinal tracking of development rather than one-time certification. This is particularly relevant for professional military education and command preparation pipelines, where leaders must demonstrate growth across increasingly complex roles.

Importantly, the framework positions assessment as a developmental asset rather than a gatekeeping mechanism. By capturing how leaders reason, reframe problems, and communicate intent, assessment outputs can be used to tailor feedback, coaching, and experiential learning. This aligns with USAF emphasis on continuous development and adaptive performance rather than static qualification.

6.1 Embedding Cognitive Readiness Sensing Within Operational Training Pipelines

A particularly important implication of the proposed framework is that assessment becomes embedded directly within operational learning pipelines rather than remaining separate from training delivery. In traditional models, evaluation typically occurs episodically through examinations, performance reviews, or isolated exercises. Such approaches provide limited visibility into how decision-making capabilities evolve over time and often fail to generate actionable developmental insight.

In contrast, the proposed agentic AI framework enables continuous measurement throughout the learning lifecycle. Scenario-based assessments can be integrated into accession programs, technical schools, professional military education, mission qualification training, and advanced leadership development. Because the assessment interactions are adaptive and conversational, evidence collection occurs naturally within realistic training activities rather than through detached testing events.

This integration enables several operational advantages for training commands and learning leaders. First, instructors gain visibility into process-level indicators of reasoning, adaptability, and judgment rather than relying solely on observable task completion. Second, learners receive immediate developmental feedback linked to specific cognitive and decision-making behaviors. Third, commanders and institutional leaders gain aggregated readiness analytics that identify systemic capability gaps across cohorts, units, and pipelines. This creates force-wide readiness visibility into adaptive warfighting capability across the learning enterprise.

Over time, this architecture supports the emergence of a continuously updated cognitive readiness profile for each Airman. Bayesian proficiency estimation enables uncertainty-aware tracking of growth across increasingly complex operational contexts, allowing developmental pathways to adapt dynamically as evidence accumulates. This is particularly valuable in AI-enabled environments where static certifications rapidly lose relevance and operational adaptability becomes the enduring competitive advantage.

6.2 Navigating AI-Enabled Decision Environments

As AI systems become embedded in operational planning, data analysis, and logistics, leadership increasingly involves managing human–AI interaction. Leaders must determine when to rely on algorithmic recommendations, when to question them, and how to integrate machine-generated insights with human judgment and values.

The proposed assessment framework explicitly incorporates this reality by distinguishing between human creative judgment and AI-mediated fluency. Distribution-aware scoring mitigates the risk that Airmen who rely uncritically on AI-generated outputs are misclassified as highly creative, while Airmen who demonstrate the ability to generate non-obvious options, challenge flawed assumptions, and adapt decisions under pressure may be undervalued. This distinction is essential for maintaining operational resilience in environments where algorithmic monoculture and convergence can undermine adaptability (Kleinberg & Raghavan, 2021; Wu et al., 2025) and Airmen capability to generate mission-viable alternatives when standard operating procedures break down.

6.3 Transferability to Civilian Workforce and Talent Systems

Although anchored in the USAF context, the framework is intentionally designed to generalize to civilian workforce systems facing similar complexity. Leaders in energy, healthcare, finance, technology, and critical infrastructure must navigate regulatory constraints, rapid technological change, and ethical tradeoffs while coordinating action across diverse stakeholders.

In these settings, traditional hiring and development practices are increasingly misaligned with AI-mediated work. Self-report measures and artifact reviews struggle to distinguish underlying capability from AI-enabled output. The distribution-aware, scenario-based approach described here provides a principled alternative, enabling organizations to measure creativity in ways that remain robust under generative AI (Doshi & Hauser, 2024; Raghavan, 2025).

This has direct implications for skills-forward leadership development, and succession planning. By separating AI fluency from creative judgment, organizations can design clearer development pathways and reduce the risk of workforce homogenization.

6.4 Learning Engineering as an Enabler of Scalable Impact

A central contribution of this work is demonstrating how learning engineering can translate theory into operational systems that scale. Rather than treating assessment, training, and analytics as separate functions, the framework integrates them within a coherent design cycle in which data from assessment informs both individual development and system-level improvement (Baker et al., 2022; Craig et al., 2025).

This integration is particularly important in defense contexts, where resources are constrained and the cost of misalignment is high. By embedding assessment within training pipelines and readiness analytics, the framework supports evidence-based decision-making without over-reliance on intuition or retrospective judgment.

6.5 Limitations and Future Directions

While promising, the framework has limitations. Scenario-based assessment necessarily abstracts from real-world complexity, and continued validation is required to strengthen links between assessment performance and operational outcomes. Additionally, governance mechanisms must evolve alongside the technology to ensure transparency, fairness, and appropriate use.

Future research will focus on expanding scenario libraries, refining distribution-aware metrics, and examining long-term predictive validity across diverse operational contexts. Further work is also needed to explore how leaders learn to use AI as a creative scaffold rather than a substitute, a distinction that has important implications for both readiness and workforce resilience (Zhang et al., 2024).

7. Discussion

This paper advances a performance-based approach to measuring mission-ready leadership that responds to three converging forces: the operational demands of modern defense environments, the limitations of traditional assessments, and the growing influence of generative AI on how human capability is expressed and evaluated. By anchoring creative thinking in USAF competency definitions while leveraging advances in learning engineering and AI-enabled assessment, the framework reframes what it means to measure readiness in an era of technological acceleration.

7.1 Reframing Creativity as Enacted Capability

A central contribution of this work is conceptual rather than purely technical. Creative thinking is treated not as latent traits inferred from self-report or reputation, but as

enacted capabilities observable through behavior in context. This framing aligns closely with USAF doctrine, which emphasizes inspiring others, organizing coordinated action, and generating new insights in novel situations (Department of the Air Force, 2025). By using scenario-based, multi-turn tasks, the framework captures creative thinking as a dynamic process that unfolds over time. This enables assessment of judgment, adaptability, and reasoning—dimensions that are critical to mission success but difficult to observe through traditional instruments. Prior work in performance-based assessment demonstrates the feasibility and value of this approach for complex skills, including collaboration and creative problem solving (Rosen, 2015; Rosen et al., 2023). The present work extends that foundation into creative thinking assessment under generative AI conditions.

7.2 Implications of Generative AI for Leadership Assessment

Generative AI fundamentally alters the signal environment for leadership and creativity assessment. As AI systems increasingly generate fluent analyses, plans, and recommendations, output quality alone becomes an unreliable indicator of human capability. Without careful design, assessment systems risk conflating AI-enabled fluency with underlying judgment, thereby misclassifying leaders and reinforcing convergence.

This paper contributes to emerging scholarship by operationalizing creativity as a distributional property rather than a one-off outcome. Distribution-aware, competitive scoring addresses the well-documented tendency of generative models to collapse toward dominant modes (Holtzman et al., 2020; Li et al., 2024; Wu et al., 2025) and aligns with recent evidence that AI can enhance individual output while reducing collective diversity (Doshi & Hauser, 2024). By evaluating distinctiveness, variability, and robustness under novelty incentives, the framework preserves meaningful signals of human creative judgment even in AI-mediated contexts (Kleinberg & Raghavan, 2021; Raghavan, 2025).

For the USAF, this distinction is not merely academic. Leaders who uncritically accept algorithmic recommendations may appear effective in routine contexts while failing under adversarial adaptation. Conversely, leaders who demonstrate disciplined creativity—questioning assumptions, reframing problems, and integrating ethical considerations—are better positioned to sustain mission advantage.

7.3 Learning Engineering as a Unifying Learning Framework

Another key contribution is methodological. The assessment framework exemplifies how learning engineering can integrate theory, design, measurement, and iteration within a coherent system (Baker et al., 2022). Rather than optimizing isolated components, the framework treats assessment as part of a broader socio-technical system that includes training pipelines, development pathways, and readiness analytics.

The use of a nested learning engineering cycle enables continuous refinement based on empirical evidence while maintaining alignment with operational goals (Craig et al., 2025). This is particularly important in defense contexts, where assessment systems must balance rigor, scalability, and accountability. The human–AI hybrid scoring architecture further reinforces this balance, ensuring that automation supports rather than replaces expert judgment.

7.4 Limitations and Open Questions

Several limitations warrant discussion. Scenario-based assessments necessarily abstract from real-world complexity, and continued validation is required to strengthen links between assessment performance and operational outcomes. Additionally, while distribution-aware metrics mitigate AI-driven homogenization, they do not eliminate all risks associated with misuse or overreliance on automated systems.

Open questions remain regarding how leaders learn to integrate AI as a creative scaffold rather than a substitute, and how assessment feedback can best support that development. Longitudinal research is needed to examine how creative thinking evolves over time and how assessment-informed interventions influence readiness at scale (Zhang et al., 2024).

8. Conclusion

Anchored in the U.S. Air Force competency definitions (Department of the Air Force, 2025 and aligned with the AETC Lines of Effort supporting the 2026 National Defense Strategy (Air Education and Training Command, 2026), the proposed AI-enabled assessment framework operationalizes mission-ready creative thinking as enacted behavior rather than inferred potential. By combining scenario-based, agentic AI task design with distribution-aware scoring and human–AI hybrid evaluation, the framework captures how individuals reason, adapt, communicate, and exercise judgment in conditions that mirror operational reality. In doing so, it addresses a growing challenge in the generative AI era: distinguishing human creative judgment from AI-enabled fluency in ways that remain fair, scalable, and predictive of real-world performance.

The framework also demonstrates how learning engineering can serve as a unifying methodology for designing assessment systems that integrate theory, measurement, and application (Baker et al., 2022; Craig et al., 2025). Rather than treating assessment as a static instrument, the approach embeds measurement within iterative design cycles that support development, readiness analytics, and organizational learning. This is particularly critical in defense contexts, where accountability, transparency, and adaptability are paramount. This paper contributes a practical, scientifically grounded model for assessing creative thinking at scale—one that aligns with mission readiness requirements while

remaining adaptable to broader workforce challenges. By making human capability visible, measurable, and developable, AI-enabled assessment can strengthen both national security and organizational resilience, ensuring that innovation on the frontlines remains guided by human judgment, values, and purpose.

Beyond the USAF, the implications of this work extend to civilian workforce and talent systems confronting similar dynamics. As generative AI becomes embedded in knowledge work, organizations across sectors face rising risks of homogenization, misattribution of capability, and erosion of differentiated judgment (Doshi & Hauser, 2024; Kleinberg & Raghavan, 2021; Wu et al., 2025). The distribution-aware, performance-based approach articulated here offers a principled path forward for measuring and developing creativity in AI-mediated environments.

In an era defined by AI-enabled warfare, contested information environments, and rapidly adapting adversaries, the strategic advantage of the United States military will depend increasingly on the quality of human judgment under pressure. Technical superiority alone is insufficient if warfighters cannot reframe problems, generate mission-viable alternatives, and adapt faster than adversaries when standard procedures fail. The framework proposed in this paper positions agentic AI assessment as foundational infrastructure for developing and sustaining cognitive readiness and adaptive warfighting capability across the Joint Force. By embedding continuous, performance-based assessment within military training pipelines, the Department of the Air Force and the broader Joint Force can move toward a future in which readiness is continuously observed, strengthened, and operationalized at scale.

References

Air Education and Training Command. (2026, April 27). *AETC lines of effort power the force behind the 2026 National Defense Strategy*. U.S. Air Force. https://www.aetc.af.mil/News/Article-Display/Article/4471421/aetc-lines-of-effort-power-the-force-behind-the-2026-national-defense-strategy/

Acar, S. (2025). Creativity assessment, research, and practice in the age of artificial intelligence. *Creativity Research Journal, 37*(2), 181-187.

Amabile, T. M., & Pratt, M. G. (2016). The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in Organizational Behavior*, *36*, 157–183.

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. Technology, Mind, and Behavior, 3(1).

Department of the Air Force (2025). *Air Force Handbook 36-2647, Competency Modeling*. Secretary of the Air Force.

Craig, S. D., Avancha, K., Malhotra, P., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training environment. In *ICICLE 2024 Conference Proceedings* (pp. 115–132).

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces collective diversity. *Science Advances, 10*(28).

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *Proceedings of ICLR 2020*.

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences, 118*(22).

Li, J., Zheng, Y., Zhou, D., et al. (2024). On the diversity collapse of large language models. In *Proceedings of NeurIPS 2024*.

National Academies of Sciences, Engineering, and Medicine (2024). *Adult Learning in the Military Context*. Washington, DC: The National Academies Press.

Raghavan, M. (2025). Competition and diversity in generative AI. *arXiv preprint arXiv:2412.08610v2*.

Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of a human-to-agent approach. *International Journal of Artificial Intelligence in Education, 25*(3), 380–406.

Rosen, Y., Jaeger, G., Newstadt, M., Bakken, S., Rushkin, I., Dawood, M., & Purifoy, C. (2023). A multidimensional approach for enhancing and measuring creative thinking and cognitive skills. *International Journal of Information and Learning Technology, 40*(4), 334–352.

Rosen, Y., & Rushkin, I. (2025). *Ignis AI PowerSkillsAssessment™: Advances in Science and AI Technology Enable Measurement of Human Power Skills*. Chestnut Hill, MA.

Rosen, Y., & Rushkin, I. (2026a). *Measuring Creativity in the Age of Generative AI: Distinguishing Human and AI-Generated Creative Performance in Hiring and Talent Systems*. Paper Presented at the 2026 BIG.AI@MIT Conference, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.

Rosen, Y., & Rushkin, I. (2026b). *From Measurement to Action: A Learning Engineering Approach to AI-Powered Assessment for Human Power Skills Development*. Paper Presented at the 2026 Learning Engineering Research Network Convening, Learning Engineering Institute, Arizona State University, Tempe, AZ.

Rosen, Y., Rushkin, I., Ang, A., Munson, L., Lopez, G., & Tingley, D. (2018). *The effects of adaptive learning in a massive open online course on learners' skill development*. Proceedings of the Fifth ACM Conference on Learning @ Scale. London, UK.

Rosen, Y., Stoeffler, K., & Simmering, V. (2020). Imagine: Design for creative thinking, learning, and assessment in schools. *Journal of Intelligence, 8*(2), 18.

Rosing, K., Frese, M., & Bausch, A. (2011). Explaining the heterogeneity of the leadership-innovation relationship: Ambidextrous leadership. *The Leadership Quarterly, 22*(5), 956-974.

Runco, M. A. (2023). AI can only produce artificial creativity. Journal of Creativity, 33(3).

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66–75.

Srivastava, A., Li, C., & Hashimoto, T. (2023). Measuring diversity in generated text using topic and distributional metrics. *Transactions of the Association for Computational Linguistics, 11*, 1234–1249.

U.S. Army Research Institute (2023). *Army Talent Attribute Framework – FY24 Annual Update Using a Mixed Methods Research Design*. United States Army Research Institute for the Behavioral and Social Sciences.

Vinchon, F., Lubart, T., Bartolotta, S., Gironnay, V., Botella, M., Bourgeois-Bougrine, S., & Gaggioli, A. (2023). Artificial intelligence & creativity: A manifesto for collaboration. *The Journal of Creative Behavior, 57*(4), 472–484.

Wingström, R., Hautala, J., & Lundman, R. (2024). Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists. *Creativity Research Journal, 36*(2), 177-193.

Wu, F., Black, E., & Chandrasekaran, V. (2025). Generative monoculture in large language models. In *Proceedings of ICLR 2025*.

Zhang, Y., Müller, M., Wang, D., et al. (2024). Beyond the prompt: How humans strategically use AI systems for creative work. In *Proceedings of CHI '24*.