Agentic AI Assessment of Creative Thinking for the U.S. Air Force

Yigal Rosen, PhD and Ilia Rushkin, PhD

Abstract

Grounded in Computational Psychometrics and Learning Engineering, this work proposes psychometrically validated, agentic AI assessment framework for measuring mission-ready creative thinking competencies in the context of talent development and training programs in the United States Air Force (USAF), with particular emphasis on complex and uncertain operational environments. While USAF competency models identify creative thinking as foundational to readiness (Air Force, 2025), prevailing assessment approaches rely largely on self-report instruments or narrowly scoped simulations. Building on the Ignis AI PowerSkillsAssessment™ framework and prior research on large-scale assessment of creative thinking (Rosen, & Ruskin, 2026; Rosen et al., 2023), this paper introduces an agentic AI–driven, scenario-based method for evaluating creative thinking competencies at scale.

Theoretical and Empirical Foundation

AI-Powered, Scenario-Based Assessment: Airmen engage with mission-relevant vignettes that require problem reframing, generation of non-obvious alternatives, questioning of assumptions, and adaptive response to new constraints, and explanation of reasoning. Responses are evaluated using a human validated AI-enabled scoring approach grounded in computational psychometrics and learning engineering principles (Rosen, & Rushkin, 2026b) to support reliability, transparency, and fairness.

Creativity skills as behavioural indicators: Creative thinking can be assessed through generation, evaluation, and improvement of ideas in realistic contexts using rigorous scoring and psychometric models.

Bayesian, Adaptive Measurement: Adaptive learning research established Bayesian proficiency estimation as a way to model evolving skill mastery with uncertainty, enabling scalable, evidence-based insights.

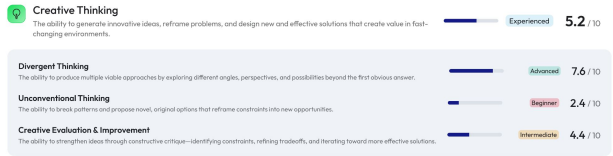
Learning Engineering in Action: The Ignis AI PowerSkillsAssessment™ is developed through an iterative learning engineering cycle that connects research insights to real-world implementation—using pilot data to continuously refine task design, scoring, proficiency models, and deployment decisions to balance rigor, fairness, and usability at scale.

Method

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.

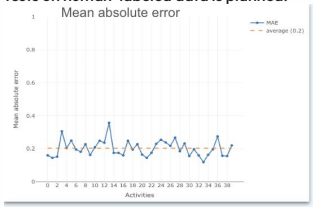
Method Cont.

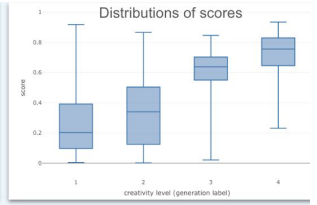
The two measures—novelty and entropy—are then combined into a single creativity score by means of a mathematical model that allows tunable meta-parameters. Importantly, this approach does not rely on opaque generative processes or subjective scoring. Instead, it provides a transparent and consistent 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.

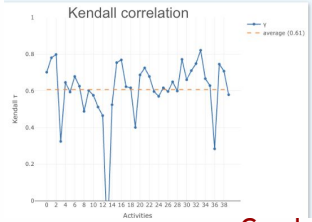


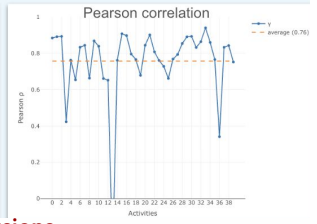
Results

Preliminary results demonstrate that creativity can be measured reliably across assessment activities. 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 is planned.









Conclusions

This research addresses the challenge of measuring creative thinking by proposing a psychometrically validated, agentic AI assessment framework for measuring mission-ready competencies, anchored in the USAF context but designed to generalize across military contexts and talent development systems facing comparable complexity. Results indicate that AI-powered, performance-based assessment of creative thinking competence is feasible, psychometrically sound, and practically useful for talent development. Ongoing work focuses on longitudinal change, predictive validity linking skills to real-world outcomes.

2026 Artificial Intelligence Workshop: Accelerating Data and Analytics Capabilities for AI
Military Operations Research Society (MORS), Carnegie Mellon University

yigal@ignisai.ai
ilia@ignisai.ai