From Measurement to Action: A Learning Engineering Approach to AI-Powered Assessment for Human Power Skills Development

Rosen, Y. & Rushkin, I.

AI-Enabled Assessment

Power Skills

Talent Development

As AI systems transform work at scale, human "power skills"—including creative thinking, communication, collaboration, leadership, analytical reasoning, productivity, and AI fluency—are increasingly critical to employability and long-term success. Yet these skills remain difficult to measure rigorously and to develop systematically. This paper is explicitly grounded in Learning Engineering and demonstrates how a nested Learning Engineering cycle—spanning challenge definition, creation, implementation, and investigation—can be embedded within a novel AI-enabled assessment system. Building on prior large-scale assessments of collaborative problem solving and creative thinking, adaptive learning research, and agentic AI, the Ignis AI PowerSkillsAssessment™ integrates evidence-centered design, generative AI, Bayesian proficiency estimation, and fairness-aware modeling into a unified operational framework. We describe the construct model, task design, scoring architecture, and early pilot results with professionals and *managers, illustrating how learning engineering can translate insights from cognitive science and psychometrics into scalable, actionable tools for talent development in the AI era.*

Introduction

The accelerating deployment of AI systems in workplaces is reshaping the division of labor between humans and machines. Routine and narrowly defined tasks are increasingly automated, while enduring value is shifting toward human capabilities that complement rather than compete with AI: creative thinking, communication, collaboration, leadership, analytical reasoning, productivity, and the ability to work effectively with AI tools. International frameworks, such as the OECD's recent Skills Outlook 2025 and the World Economic Forum's Future of Jobs Report 2025, converge on the view that human power skills—such as creative thinking, adaptability, communication, and leadership—are central to future employability, lifelong learning, and innovation across sectors (OECD, 2025; World Economic Forum, 2025). Despite this recognition, most existing tools for assessing these capabilities rely on self-report questionnaires or small-scale divergent-thinking tests, which suffer from limited construct validity, social desirability bias, and poor alignment with real-world performance. There is a pressing need for assessment systems that can (a) capture how people actually perform in realistic, cognitively demanding situations, (b) provide interpretable, actionable evidence to support learning and development, and (c) scale across contexts, roles, and industries. The Ignis AI PowerSkillsAssessment™ was designed to address this need by integrating psychometrically grounded constructs, AI-mediated task environments, automated scoring, and Bayesian proficiency estimation into a unified system. This paper describes how learning-engineering principles guided the translation of research on power skills into an operational assessment platform that can support scalable and equitable talent development.

This work is situated within the field of Learning Engineering, which integrates learning sciences, human-centered design, data, and engineering practices to translate research insights into scalable, evidence-based learning systems. As articulated by Baker et al. (2022) and Goodell et al. (2023), Learning Engineering emphasizes iterative design, implementation, and evaluation grounded in real-world constraints. The Ignis AI PowerSkillsAssessment™ represents a concrete instantiation of this approach, operationalizing a nested Learning Engineering cycle within a broader AI-enabled platform to support the measurement and development of complex human skills (Thai et al., 2023).

Theoretical and Empirical Foundations

The design of the PowerSkillsAssessment™ draws on three interlocking strands of prior work. First, research on large-scale assessment of "hard-to-measure" social–cognitive skills, including the PISA 2015 Collaborative Problem Solving assessment, demonstrated that computer-driven conversational agents can emulate key aspects of human collaboration—such as establishing shared understanding, managing conflict, and coordinating action—within standardized digital environments (OECD, 2017; Rosen, 2014, 2015). Results showed that performance in human-to-agent settings correlated strongly with human-to-human collaboration while providing richer and more consistent behavioral data. Second, the PISA 2022 Creative Thinking assessment offered the first internationally validated framework and computer-based assessment for creative thinking in 15-year-olds, conceptualizing creative thinking as the ability to generate, evaluate, and improve ideas across multiple domains (OECD, 2024; Rosen, Stoeffler, & Simmering, 2020). This work moved beyond small-scale divergent-thinking tasks and self-report measures by using open-ended, scenario-based activities evaluated through rigorous scoring rubrics and psychometric modeling. Third, research on adaptive learning and proficiency estimation at Harvard and in large-scale online courses provided a foundation for Bayesian modeling of skill mastery and dynamic, evidence-based feedback. Projects such as the Adaptive Learning Open Source Initiative (ALOSI) applied Bayesian knowledge-tracing and open adaptive engines in MOOCs to estimate learners' evolving mastery states across multiple skills and adjust instruction accordingly (Rosen et al., 2017, 2018). These strands—agentic AI in assessment, creative thinking measurement at scale, and Bayesian adaptive modeling—collectively informed the design of a power-skills assessment architecture that is both scientifically robust and operationally scalable.

Power Skills Framework and Construct Model

The system begins with a construct architecture that defines each skill through observable, evidence-aligned components. A construct model displaying major power skills—Communication, Collaboration, Creative Thinking, Leadership, Analytical Reasoning, Productivity, and AI Fluency. Each domain is further decomposed into subskills grounded in prior research and psychometric theory. For example, creative thinking is operationalized in terms of originality, flexibility, elaboration, and idea improvement, consistent with the PISA 2022 framework and creativity literature (Rosen et al., 2020; Kaufman & Beghetto, 2009). Communication includes clarity, audience awareness, structure, and tone; collaboration includes coordination, perspective-taking, and constructive feedback; leadership focuses on direction-setting, alignment, and support; analytical reasoning incorporates interpretation, inference, and argumentation; productivity includes goal-setting, prioritization, and follow-through; and AI fluency focuses on problem formulation, prompt strategy, and critical evaluation of AI-generated outputs. This construct model guides all downstream design decisions: item specifications, task environments, rating rubrics, scoring models, and reporting structures. It also provides a basis for evaluating construct validity by examining whether observed performance patterns align with theoretically expected relationships among subskills and domains (Messick, 1989; Mislevy, 2013; Totino & Kessler, 2024).

Task Design and AI-Assisted Content Generation

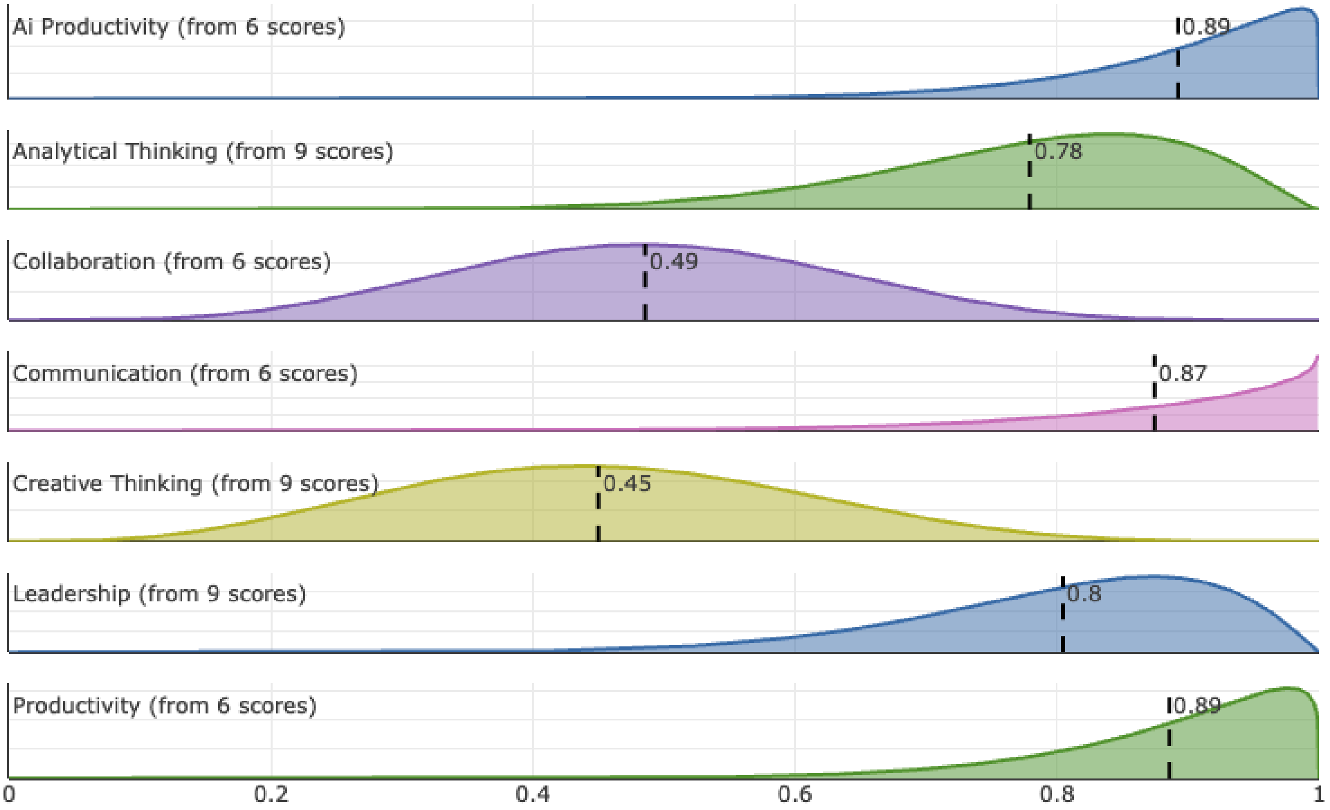
The assessment uses open-ended and scenario-based tasks that emulate realistic cognitive and interpersonal demands. Examples include: (a) Composing a response to a complex stakeholder email (Communication + Leadership); (b) Providing structured peer feedback on a teammate's proposal (Collaboration + Communication); (c) Generating multiple solutions to an ambiguous workplace challenge and refining one into a detailed plan (Creative Thinking + Analytical Reasoning); and (d) Evaluating AI-generated outputs for a given prompt, identifying limitations, and revising the prompt to improve the result (AI Fluency + Productivity). Task development follows evidence-centered design principles (Mislevy et al., 2003): for each subskill, designers specify the claim (what is being inferred), the evidence (what behaviors indicate proficiency), and the task features (what situations are likely to elicit that evidence). This structure ensures alignment between intended constructs and observable behaviors. To accelerate content development while preserving validity, the system uses a controlled generative-AI pipeline. Large language models are prompted with structured templates and constraints derived from the construct model and item specifications. Candidate tasks are generated, automatically screened for basic quality and redundancy, and then reviewed by human experts for construct alignment, clarity, and fairness. This approach allows coverage of multiple industries and roles while keeping humans in the loop for high-value decisions, such as ensuring that tasks truly elicit targeted subskills and are culturally and contextually appropriate.

Scoring Architecture and Bayesian Proficiency Estimation

Responses are scored using a hybrid human–machine approach. For each task, a rubric defines performance levels for relevant subskills, with behavioral descriptors anchored in prior research (e.g., on creative thinking levels, collaborative problem-solving processes, and written communication quality). A portion of responses is double-scored by trained human raters to establish reliability and calibrate automated scorers. Automated scoring models leverage large language models fine-tuned or prompted with rubric exemplars. To mitigate construct-irrelevant variance and bias, prompt templates include rubric descriptors, and scoring outputs are audited for consistency and fairness across subgroups. Skill proficiency is estimated within a Bayesian latent-trait framework (e.g., multidimensional IRT or related models; von Davier, 2010). For each participant and each skill, the model maintains a posterior probability distribution over proficiency on a 0–1 scale. As responses accumulate across tasks, the posterior updates, narrowing as evidence increases. Figure 1 illustrates an example posterior distribution.

Figure. 1.

Illustrative example for power skills estimation for an individual participant



The plots show the posterior probability distributions of the participant's proficiency level on each skill, measured on the scale from 0 to 1. The dashed vertical lines with numeric values indicate the means of those distributions, which serve as the reportable single-value estimates of proficiency. This probabilistic model aligns with learning-engineering principles by supporting both assessments (what is the current skill level?) and learning optimization (what additional evidence is most useful?).

Preliminary Empirical Evidence

Pilot studies with professionals and managers were conducted to evaluate the psychometric properties of the Ignis AI PowerSkillsAssessment™ and to illustrate how Learning Engineering connects design decisions to data-driven improvement across iterative cycles. Consistent with a nested Learning Engineering methodology, these pilots were used not only to validate measurement quality, but also to inform refinement of task design, scoring models, and proficiency estimation (Study I and Study II). Together, the studies focused on instrument reliability, factor structure, item performance, uncertainty modeling, and fairness across diverse populations.

Study I was conducted in a low-stakes setting with a diverse U.S. adult sample recruited via the Prolific platform (n = 353). Participants completed scenario-based tasks spanning creative thinking, communication, collaboration, leadership, analytical reasoning, productivity, and AI fluency. Psychometric analyses indicated strong internal consistency across domains, with Cronbach's α ranging from .82 to .91—well within accepted ranges for short, performance-based assessments involving open-ended and ordered multiple-choice responses. Exploratory and confirmatory factor analyses supported the hypothesized multidimensional structure of the Power Skills framework, with each domain loading strongly on its intended latent factor and moderate inter-factor correlations, consistent with related but distinct constructs. Item-level analyses demonstrated a healthy spread of difficulty and discrimination parameters, and standard workforce item-selection criteria resulted in a 96% acceptance rate in this preliminary pilot. Qualitative inspection of constructed responses further revealed meaningful differences between higher- and lower-performing participants, particularly for indicators of originality, adaptability, and integrative decision-making in creative thinking and leadership tasks.

Study II extended this work in a higher-stakes, applied context with employees from two technology organizations (n = 36), representing roles in product management, software development, UX/UI design, and senior leadership. This pilot included an expanded task set and was designed to evaluate psychometric calibration, participant experience, and the behavior of the Bayesian proficiency estimation model under real-world conditions. AI-assisted scoring models, supported by human-in-the-loop validation, were applied to constructed and ordered responses, and posterior proficiency distributions were generated for each skill. Results indicated robust reliability across domains (average α = .85), and confirmatory factor analyses again supported the framework's skills and subskills structure. Item information curves showed optimal precision in mid- to high-proficiency ranges, which is particularly important for identifying emerging leaders and supporting targeted development. Approximately 75% of items met workforce-grade psychometric criteria in this applied setting, providing clear guidance for ongoing optimization of the assessment

Across both studies, Bayesian proficiency estimation enabled explicit modeling of uncertainty, with posterior distributions narrowing as additional evidence was incorporated through more items, multi-tagged tasks, and supplemental data sources. This approach allowed proficiency scores to be reported alongside credible intervals, reinforcing interpretability and appropriate use of results. Convergent validity analyses comparing performance-based estimates with self-report and peer-review indicators showed moderate correlations, consistent with prior research suggesting that behavioral assessments capture related but non-redundant variance compared to perception-based measures. Collectively, these findings provide promising preliminary evidence that the Ignis AI PowerSkillsAssessment™ can generate reliable, fair, and developmentally informative measures of human power skills in both low- and higher-stakes contexts, while exemplifying how Learning Engineering principles support continuous improvement from empirical evidence to system refinement.

Learning Engineering in Action: From Insights to Implementation

The design and refinement of the PowerSkillsAssessment™ follow a learning engineering cycle: (a) Insight generation from prior research and large-scale implementations (e.g., PISA 2015/2022, adaptive learning in MOOCs) informs the construct model and task design; (b) Prototyping of tasks, scoring prompts, and proficiency models is carried out in small-scale pilots; (c) Data collection and analysis through pilots provide empirical evidence on reliability, validity, item performance, fairness, and user experience; (d) System redesign uses this evidence to revise tasks, rubrics, scoring procedures, and proficiency estimation, including the refinement of AI prompts and filters; and (e) Implementation in real-world contexts (e.g., workforce development programs, higher education settings) produces additional data to inform both product evolution and research questions.

This iterative process exemplifies learning engineering: it integrates educational science, design, AI, and data systems to produce tools that are not only technically sophisticated but also usable, interpretable, and aligned with real-world constraints. Insights from pilots directly affect implementation decisions, such as how many tasks per skill are required to achieve acceptable uncertainty thresholds, which task types best differentiate proficiency levels, and how to balance efficiency with depth of evidence.

Discussion and Future Directions

The early evidence suggests that an AI-powered, performance-based assessment of human power skills is technically feasible, psychometrically defensible, and practically applicable in talent development contexts. The combination of construct modeling, AI-assisted task generation, rubric-grounded scoring, Bayesian proficiency estimation, and fairness auditing constitutes a coherent learning-engineering pipeline.

Several directions for future work are particularly promising. First, longitudinal studies are needed to evaluate how proficiency estimates change over time in response to targeted learning experiences, including AI-supported coaching and practice. Second, predictive validity should be examined by linking power skills profiles to real-world outcomes such as talent development, job performance, promotion, and engagement. Third, the interaction between AI fluency and other power skills merits deeper investigation: for example, how individuals who effectively leverage AI tools differ in their creative output and collaboration quality compared to those who do not.

From a learning engineering perspective, continued work is required to refine the feedback layer—how to transform posterior distributions and profile patterns into actionable, personalized recommendations for learners, coaches, and organizations. This includes experimentation with different feedback formats, visualizations, and integration points within learning and talent platforms.

References

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). https://doi.org/10.1037/tmb0000058

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: Solving for Complexity at Scale (pp. 115–132). https://doi.org/10.59668/2109.21735

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. Journal of Military Learning, Conference Edition. https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/

Goodell, J., & Kolodner, J. (Eds.). (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge.

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four C model of creativity. Review of General Psychology, 13(1), 1–12.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York, NY: Macmillan.

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. Military Medicine, 178(10), 107–114.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1(1), 3–62.

OECD. (2017). PISA 2015 results (Volume V): Collaborative problem solving. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2017/11/pisa-2015-results-volume-v_g1g83e07/9789264285521-en.pdf

OECD. (2024). PISA 2022 results (Volume IV): Creative thinking. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/pisa-2022-results-volume-iv_125a58b3/5a849c2a-en.pdf

OECD. (2025). OECD Skills Outlook 2025: Building the Skills of the 21st Century for All. OECD Publishing. https://www.oecd.org/en/publications/oecd-skills-outlook-2025_26163cd3-en.html?utm_source=chatgpt.com

Page, E. B. (2000). Computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 68(2), 127–142.

Rosen, Y. (2014). Comparability of conflict opportunities in human-to-human and human-to-agent online collaborative problem solving. Technology, Knowledge and Learning, 19(1–2), 147–174.

Rosen, Y. (2015). Assessing students in human-to-agent settings to inform collaborative problem-solving learning. Journal of Educational Measurement, 52(4), 345–370.

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

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., Ang, A., Fredericks, C., Tingley, D., & Blink, M.-J. (2017). Designing adaptive assessments in MOOCs. In Proceedings of the Fourth ACM Conference on Learning at Scale.

Rosen, Y., Rushkin, I., Ang, A., Munson, L., Lopez, G., Tingley, D., & Weber, G. (2018). The effects of adaptive learning in a massive open online course on learners' skill development. In Proceedings of the Fifth ACM Conference on Learning at Scale. https://dl.acm.org/doi/10.1145/3231644.3231651

Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2023). Learning engineering is human-centered. In J. Goodell (Ed.), The learning engineering toolkit (pp. 83–124). Routledge.

Totino, L., & Kessler, A. (2024). "Why did we do that?" A systematic approach to tracking decisions in the design and iteration of learning experiences. Journal of Applied Instructional Design, 13(2).

von Davier, M. (2010). Statistical models for test equating, scaling, and linking. Springer.

World Economic Forum. (2025). The Future of Jobs Report 2025. World Economic Forum. https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf?utm_source=chatgpt.com