Peer-Reviewed Research Paper Presented at the International Test Commission Conference, Testing for good: Using tests to benefit individuals and society[1], Auckland, New Zealand: July 1, 2026

Advances in Native Agentic AI Assessment of Power Skills: Implications for Measurement and Talent Development

Yigal Rosen, Ph.D[2]. and Ilia Rushkin, Ph.D. (Ignis AI, Boston, MA)

Abstract

Recent advances in artificial intelligence (AI) and computational psychometrics have enabled scalable measurement of complex workforce competencies, including creative thinking, communication, collaboration, and leadership. This paper presents a native agentic AI assessment framework grounded in measurement theory that integrates constructed-response tasks, Bayesian latent proficiency modeling, and human-in-the-loop AI scoring to measure developmental continua of power skills. Three validation studies (total n > 1,000) evaluated psychometric quality, fairness, and AI–human scoring alignment. Study 1 demonstrated strong psychometric performance using item response theory, including high item discrimination, well-distributed item difficulty, and no meaningful differential item functioning across gender, race/ethnicity, or age groups, supporting fairness and construct validity in assessments of individual contributors and managers. Study 2 replicated these findings in a leadership-focused assessment, demonstrating robustness across domains and contexts. Study 3 examined AI-assisted scoring of constructed responses and found substantial agreement between AI and human raters (QWK = 0.788), exceeding average inter-rater human agreement. Together, these findings suggest that agentic AI assessment systems can achieve psychometric rigor comparable to traditional measures while extending reliable measurement to multidimensional, real-world human capabilities relevant to education, workforce development, and leadership evaluation.

Keywords: Workforce competency measurement; Power skills assessment; Computational psychometrics; Constructed-response scoring; AI scoring; Human–AI scoring alignment.

[1] This is an adaptation from the Māori saying: Mā tōu rou, mā taku rourou ka ora te iwi. With your food basket and my food basket the people will thrive.

[2] Correspondence: yigal@ignisai.ai

1. Introduction

Artificial intelligence (AI) is transforming how organizations identify, develop, and evaluate human capabilities. Across education, workforce development, talent management, and leadership assessment, increasing attention has shifted from measuring technical knowledge alone toward assessing complex human competencies that enable individuals to perform effectively in dynamic, uncertain, and collaborative environments (National Academies of Sciences, Engineering, and Medicine, 2024). These competencies—often referred to as power skills, durable skills, or human skills—include creative thinking, communication, collaboration, analytical reasoning, leadership, adaptability, and related capabilities that are increasingly recognized as critical predictors of success across industries, roles, and contexts (Rosen & Rushkin, 2025; Rosen & Rushkin, 2026a).

Despite broad agreement regarding the importance of these competencies, their measurement remains a longstanding challenge. Traditional assessment approaches frequently rely on self-report questionnaires, supervisor ratings, interviews, or narrowly defined performance tasks. While such methods can provide useful information, they often suffer from limitations in validity, scalability, susceptibility to response biases, and insufficient sensitivity to real-world performance (Rosen, 2015a; Rosen et al., 2020). Self-report measures capture perceptions rather than demonstrated capability, whereas human ratings can be costly, difficult to standardize, and vulnerable to inconsistency. Moreover, many existing assessments were developed for educational or organizational environments that predate recent advances in generative AI, raising important questions about whether conventional approaches remain adequate for measuring complex human competencies in AI-mediated environments (Rosen & Rushkin, 2026b; Rosen & Rushkin, 2026c). The emergence of large language models (LLMs) and other generative AI technologies presents both an opportunity and a challenge for assessment science. On one hand, advances in natural language processing and computational psychometrics enable the creation of rich scenario-based assessments, automated scoring systems, and adaptive measurement frameworks capable of evaluating complex human behaviors at unprecedented scale (Hao, et al, 2024 2024; Baker et al., 2022; Craig et al., 2025; von Davier, et al., 2023). On the other hand, the growing ability of AI systems to generate sophisticated written responses complicates traditional approaches that infer human capability primarily from assessment outputs. When high-quality responses can be produced or substantially enhanced through AI assistance, distinguishing underlying human proficiency from AI-enabled fluency becomes increasingly difficult (Doshi & Hauser, 2024; Kleinberg & Raghavan, 2021). Recent research suggests that generative AI can extend assessment beyond traditional academic constructs and support measurement of complex interpersonal competencies such as communication, collaboration, negotiation, and collaborative problem solving. Unlike conventional constructed-response scoring, which typically evaluates individual written products, assessment of communication and collaboration requires interpretation of discourse acts, interaction patterns, information sharing, perspective taking, and social regulation behaviors embedded within dynamic exchanges. Emerging evidence indicates that large language models can successfully code such communication data with levels of accuracy comparable to trained human raters, creating the possibility of scalable assessment of skills that have historically been difficult and expensive to measure (Hao et al., 2026). These developments are particularly important because communication and collaboration represent core workforce competencies that are often underrepresented in large-scale assessment systems due to scoring complexity and cost.

These developments have intensified interest in performance-based assessments that evaluate how individuals reason, communicate, collaborate, and solve problems within realistic contexts rather than relying solely on static knowledge tests or self-reported dispositions. Scenario-based assessment approaches are particularly promising because they can elicit observable behaviors associated with complex competencies while preserving contextual authenticity (Rosen, 2015b). Advances in AI now

enable these assessments to be delivered through interactive, multi-turn environments that dynamically adapt to participant responses, generating rich evidence about underlying skills and decision-making processes (Rosen & Rushkin, 2025). Such approaches align with contemporary perspectives in learning engineering and assessment design, which emphasize the integration of theory, evidence collection, scoring, and validation within coherent measurement systems (Baker et al., 2022; Craig et al., 2025). At the same time, the increasing use of AI in assessment raises important psychometric and ethical questions. For AI-augmented assessments to be adopted in educational, workforce, or leadership contexts, they must demonstrate evidence of reliability, validity, fairness, and transparency comparable to—or exceeding—that of traditional measurement approaches. Concerns regarding algorithmic bias, demographic fairness, explainability, and human oversight have become central to discussions of responsible AI in assessment and talent management (National Academies of Sciences, Engineering, and Medicine, 2024). Consequently, the validity of AI-based measurement systems cannot be assumed based on technological sophistication alone; rather, it must be established through rigorous psychometric evaluation.

Recent advances in computational psychometrics provide a foundation for addressing these challenges. Bayesian proficiency modeling enables estimation of latent skill profiles while explicitly representing uncertainty (Rosen et al., 2018; von Davier et al., 2023). Item Response Theory (IRT) provides well-established methods for evaluating item characteristics and measurement precision (von Davier, 2024). Differential Item Functioning (DIF) analyses allow systematic investigation of potential bias across demographic groups (Teresi et al., 2021; Halpin, 2024). Similarly, human–AI agreement metrics offer a means of evaluating whether automated scoring systems produce judgments comparable to those of trained human raters (Hao et al., 2024; Li et al., 2025). Together, these approaches create a pathway toward scalable assessment systems that maintain psychometric rigor while leveraging the efficiency and flexibility of AI.

The present study builds upon these developments by introducing and validating an AI-augmented assessment framework designed to measure human power skills through scenario-based, constructed-response tasks. The framework integrates three key components: (a) AI-generated and expert-reviewed assessment content grounded in a hierarchical ontology of power skills, (b) human-in-the-loop AI scoring of open-ended responses, and (c) Bayesian latent proficiency estimation that synthesizes evidence across tasks into interpretable skill profiles. Unlike traditional assessments that rely primarily on selected-response items or self-reports, the framework is designed to capture observable manifestations of complex competencies through realistic workplace and leadership scenarios. Importantly, the framework conceptualizes power skills as multidimensional and developmental constructs that can be inferred from patterns of performance across diverse contexts rather than from isolated test items. This perspective is consistent with contemporary theories of competency assessment and learning engineering, which emphasize the importance of measuring applied capability in authentic situations (Rosen et al., 2023; Baker et al., 2022). It also reflects growing recognition that workforce success increasingly depends on the integration of cognitive, interpersonal, and adaptive capabilities that extend beyond traditional measures of academic achievement or technical expertise.

Importantly, agreement between AI and human raters represents only one component of a broader validity argument for AI-assisted scoring. Recent work has emphasized that AI scoring systems should also demonstrate fairness, subgroup consistency, transparency, and reliability comparable to human scoring processes (Hao et al., 2026). In particular, research examining AI coding of collaborative communication has proposed evaluating whether AI-generated judgments remain stable across demographic groups and whether AI-human agreement exhibits patterns comparable to human-human agreement. Such evidence is especially important when assessing power skills because communication

styles, leadership approaches, and collaboration behaviors may vary across demographic and cultural groups while still representing equivalent underlying competencies.

The primary purpose of this research is to evaluate the psychometric quality, fairness, and scoring validity of this AI-augmented power skills assessment. Specifically, three research questions guide the investigation:

1. What are the psychometric properties of agentic AI, scenario-based assessment items with respect to item difficulty, item discrimination, and internal consistency?
2. To what extent do the assessment items exhibit evidence of differential item functioning across protected demographic groups, including gender, race/ethnicity, and age?
3. To what extent do AI-generated scores align with human scoring judgments when evaluating constructed responses?

To address these questions, three empirical studies involving more than 1,000 participants were conducted using independently sampled cohorts. Two studies evaluated the psychometric characteristics and fairness of assessments measuring multiple power skills across professional and leadership contexts, while a third study examined agreement between AI-generated scores and trained human raters. Collectively, these studies provide evidence regarding the feasibility of using AI-augmented assessment systems for reliable and equitable measurement of complex human capabilities. The findings demonstrate that AI-generated, scenario-based assessments can achieve strong psychometric performance, including high item discrimination, appropriate difficulty distributions, and minimal evidence of demographic bias. Furthermore, AI scoring exhibited substantial agreement with human raters, achieving levels of consistency that exceeded average human-to-human agreement. These results suggest that AI-augmented assessment frameworks can extend the reach of psychometrically rigorous measurement beyond traditional testing formats, enabling scalable evaluation of complex human competencies relevant to education, workforce development, leadership assessment, and talent management. By integrating advances in artificial intelligence, computational psychometrics, and performance-based assessment, this work contributes evidence that scalable measurement of human power skills is both technically feasible and psychometrically defensible. More broadly, it provides a model for how AI can augment—rather than replace—human judgment within assessment systems designed to evaluate the capabilities that remain uniquely important in an increasingly AI-enabled world.

2. Materials and Methods

2.1. Assessment Instruments

The present study utilized the Ignis AI Power Skills Assessment™ framework, a performance-based assessment system designed to measure human power skills through scenario-based interactions with agentic AI (Rosen & Rushkin, 2025). Unlike traditional self-report inventories, the framework operationalizes power skills as observable behaviors demonstrated during realistic workplace and leadership situations. The assessment is grounded in the human power skills ontology, which defines human capabilities as multidimensional constructs that can be elicited, observed, and evaluated through authentic performance tasks (Rosen & Rushkin, 2026a; Rosen & Rushkin, 2026b; Rosen & Rushkin, 2026c). The Human Power Skills Ontology organizes competencies into seven broad domains: Creative Thinking, Leadership, Communication, Collaboration, Analytical Thinking, Productivity, and AI Fluency. Each domain is further decomposed into theoretically derived subskills that represent distinct aspects of

performance. For example, Creative Thinking includes Divergent Thinking, Unconventional Thinking, and Creative Evaluation and Improvement, while other domains are similarly represented through hierarchically organized competencies and behavioral indicators. This ontology serves as the foundation for assessment design, evidence collection, scoring, and reporting. Assessment tasks are delivered through multi-turn, scenario-based interactions with AI agents. Participants engage in realistic professional situations characterized by ambiguity, competing priorities, incomplete information, and evolving constraints. Rather than selecting responses from predefined options, participants generate open-ended responses to prompts presented by an AI agent (see Figure 1).

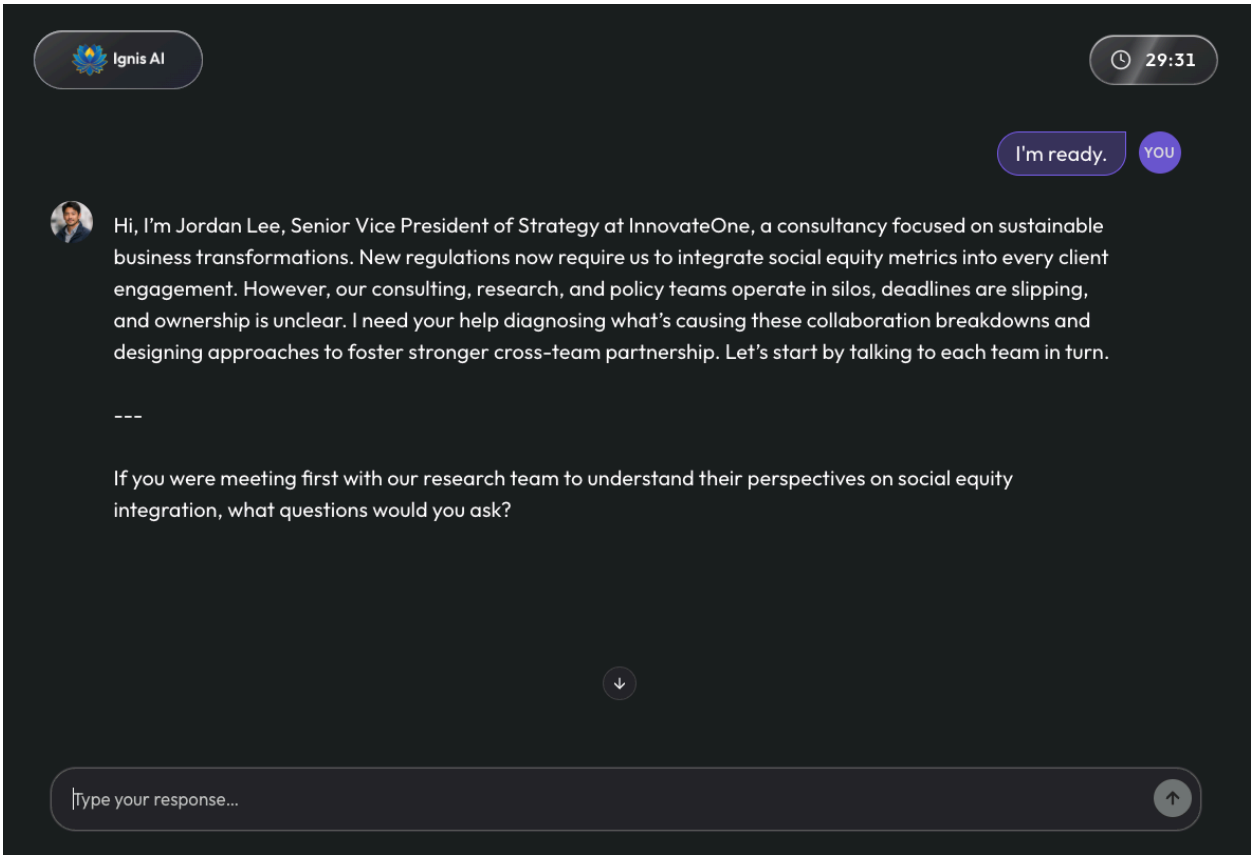

Figure 1. Example for Ignis AI PowerSkillsAssessment™ 30-minutes activity designed and validated to measure leadership and collaboration skills

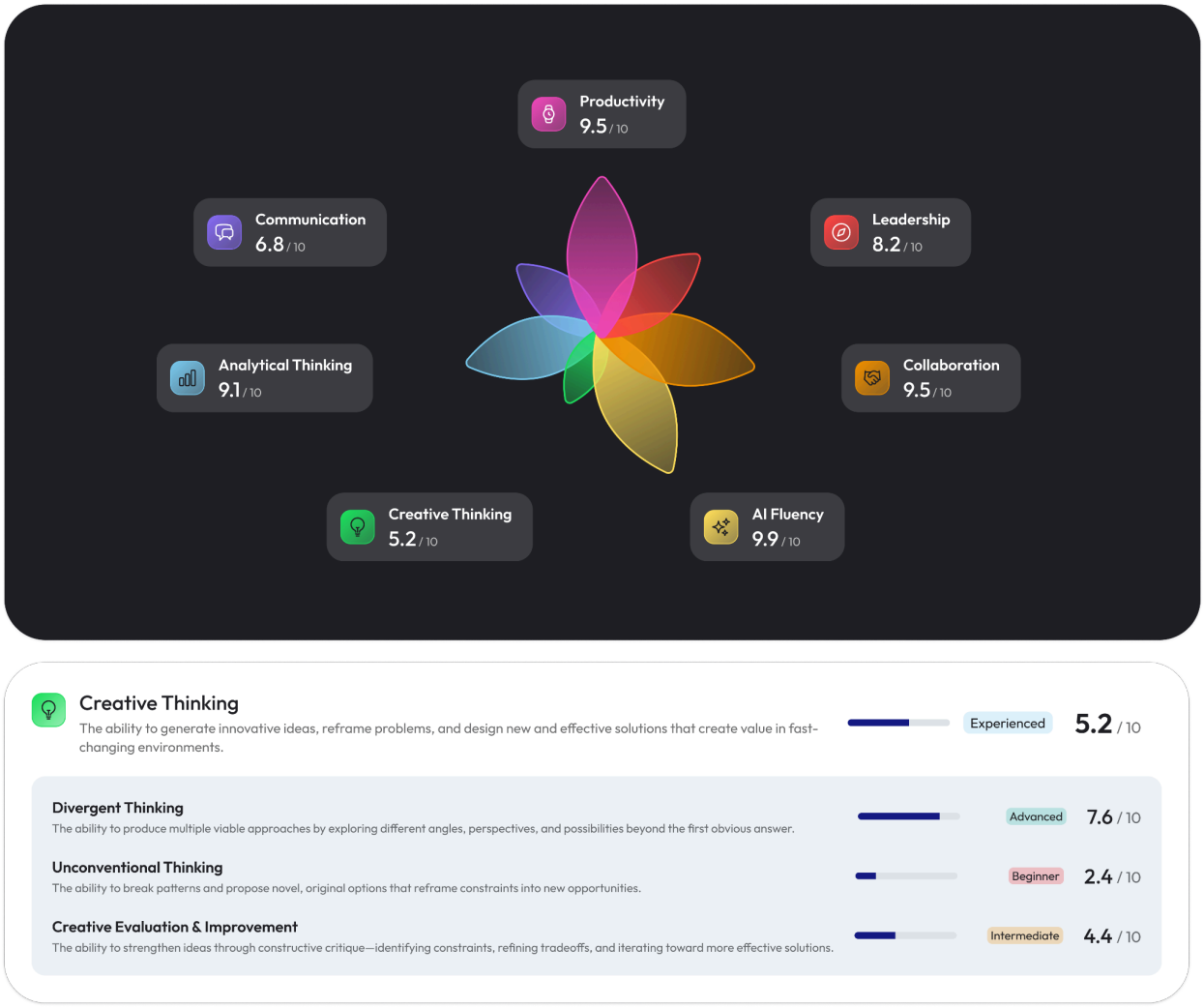

Figure 2. Ignis AI PowerSkillsPrint™ designed to provide participants with insights on their power skills proficiencies upon completion of all assessment activities

This agentic design enables the collection of process-level evidence regarding how individuals frame problems, generate alternatives, evaluate tradeoffs, communicate rationale, and adapt their thinking over time. The assessment framework follows principles of learning engineering and evidence-centered design, aligning task features with targeted competencies and intended score interpretations. Scenarios are generated using a hybrid AI–human subject matter expert (SME) process and are calibrated to specified levels of complexity. Complexity is controlled through parameters such as linguistic demands, conceptual abstraction, inferential reasoning requirements, working memory load, and the number of reasoning steps required to complete the task. This approach allows the framework to produce assessments appropriate for diverse populations and organizational contexts while maintaining alignment with the targeted construct. Responses are evaluated using a human–AI hybrid scoring architecture. Large language models and related AI systems extract evidence from participant responses and generate preliminary scores based on competency-specific scoring rubrics and machine learning models. This approach is consistent with established frameworks for the development, validation, and operational use of automated scoring systems (Williamson et al., 2012).

Human experts contribute to rubric development, score calibration, quality assurance, bias monitoring, and adjudication of ambiguous cases. This hybrid approach is intended to balance scalability with psychometric rigor and transparency.

The framework incorporates psychometric quality-control procedures, including item response theory analyses, differential item functioning analyses, and human–AI agreement studies. Previous investigations using diverse participant samples found no systematic evidence of bias across major demographic groups and demonstrated acceptable item difficulty and discrimination characteristics across assessment activities. These procedures support the validity and fairness of score interpretations derived from the assessment system. Following completion of the assessment, proficiency estimates are aggregated to produce an individualized Ignis AI PowerSkillsPrint™, a multidimensional profile representing performance across the assessed power skills and their underlying subskills (see Figure 2). Scores are reported on calibrated proficiency scales and accompanied by developmental interpretations designed to support both talent development and organizational decision-making.

2.2. Skill evaluation flow

Our skill evaluation flow consists of two distinct stages: assessment scoring and proficiency estimation. The assessment content itself is produced by a multi-step GenAI-based process that allows generating diverse versions of content to specification.

2.2.1. Assessment scoring

Assessment scoring is performed using a GenAI-based framework that evaluates the participants' responses to assessment items according to our skill framework. Generally, each item is associated, with different weights, to several skills from our framework, in which case the response is evaluated with respect to each skill, producing multiple scores (one per associated skill).

2.2.2. Skill proficiency estimation

Next, proficiency estimation is done using a Bayesian latent-variable model that ingests assessment scores as signals (other sources of signals can also be used). The model takes into account the hierarchical relations among skills in the framework and outputs posterior distribution for each required skill. The model can be applied to an individual's data or to group data (estimation of the skill profile of a cohort of participants).

The mean of the posterior distribution is a reportable single-number description of the proficiency, but full distributions are used to provide confidence intervals as well as for further practical uses, such as (but not limited to):

- Estimation of a fit for particular skill requirements that could be expressed as a set of minimum thresholds for a number of skills
- Estimation of the statistical significance and effect size of the difference before and after a talent-development intervention
- Estimation of skill profile similarities among people or cohorts, e.g., by computing the Wasserstein metric (Figalli & Glaudo, 2021) among the skill posterior distributions.

2.3 Assessment and scoring validation

2.3.1. Differential Item Functioning (DIF) analysis

Our standard implementation of the DIF analysis, which we intend for ongoing use in the product, models each item score with a fractional-binomial generalized linear model as *score ~ rest_score + group + rest_score:group*. Here the score was taken as the score on the primary-tag skill of the item, rest-score was

defined as the sum of scores the participant received on all the remaining items in the activity and group was one of the categorical demographic variables described in the above. The estimated fit coefficient for the term *group* describes the uniform DIF, and that for the term *rest_score:group* describes the non-uniform DIF. An item is marked if either of these coefficients is statistically significant ($p>p_{\text{threshold}}$), 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.

2.3.2. Psychometric item analysis

We employ a 2PL-IRT model on dichotomized scores, for each Power Skill separately, to estimate the item characteristics (difficulty and discrimination). For interpretability, both parameters are transformed after the fit to [0,1] range. For difficulty, as well as for participant ability, the transformation is the standard normal CDF, and for discrimination it is $x/(1 + x)$. Because performing such transformations is not yet common in literature, some remarks on the interpretation methodology are in order. For IRT discrimination, a common interpretation scheme, going back at least to (Baker, 2001, p. 34), uses the labels "very low," "low," "moderate," "high," very high" with cut-points 0.35, 0.65, 1.35, 1.7 in the untransformed space, meaning that the transformed values in the range [0.63, 1) are interpreted as "very high" ($0.63 \approx 1.7/(1 + 1.7)$). For transformed difficulty and participant ability one can use the same labels with cut-points 0.15, 0.4, 0.6, 0.85. These are virtually equivalent to cut-points -2, -1, 1, 2 in the untransformed space, which are motivated as the convenient integer z-values for the standard normal distribution of ability in the model.

Additionally, we keep track of standard classical test-theoretical measures (mean and standard deviation of scores, item-total correlations, Cronbach's alpha).

2.3.3. Scoring agreement with humans

We use the quadratic weighted kappa (QWK) and, in case of more than two raters, the Krippendorff's alpha to evaluate inter-rater reliability, applying it both to a group of human scorers and to evaluate the differences between those and our scoring model. Additionally, we monitor Pearson and Kendall correlations between the model score and the consensus score of the human raters.

2.3.4. Scoring fairness and subgroup consistency

Following recent recommendations for evaluating AI-assisted scoring systems, we examined subgroup consistency of assessment outcomes across major demographic groups. Whereas traditional validation studies often focus primarily on overall scoring agreement, emerging frameworks suggest that AI scoring systems should also demonstrate comparable behavior across demographic subgroups (e.g., Hao et al., 2026). Consistent with this perspective, differential item functioning analyses were conducted to evaluate whether assessment tasks functioned similarly across gender, race/ethnicity, and age groups. In addition, human–AI scoring agreement was examined within demographic subgroups to evaluate whether scoring consistency differed systematically across populations. These analyses provide complementary evidence regarding fairness and responsible deployment of AI-assisted assessment systems.

2.4 Description of Studies

Several studies were performed on the crowd-sourcing platform Prolific. In two of them (Studies I and II), conducted in March-April 2026, participants completed several full assessments scenarios each: one scenario per a Power Skill as its primary skill association. There was no participant overlap between the two studies. Two editions of content were used: Professional Edition content (with 7 Power Skills) in Study 1 and Leadership Edition content (with 3 Power Skills) in Study 2. All items in the scenarios were free-response questions, generated using our AI-based framework. The participant cohorts were screened to USA or UK as the country of residence, and balanced on several demographic variables: sex/gender (Male vs. Female), ethnicity (Asian/Black/White), age group (18-40 yo vs. 41-69 yo), employment status (binary) or student status (binary). In Study 2 the cohort was further screened to those participants whose most recent work role was identified as sufficiently senior: CEO or C-suite executive, president, vice president, director or associate director, senior manager, or manager.

In Study 1, the number of participants was n = 613, the number of assessment items was 54, distributed across 7 Power Skills of Professional Edition. In Study 2, the number of participants was n = 431, the number of items was 40, distributed across 3 Power Skills of Leadership Edition.

Another study (Study 3) was conducted on Prolific in April 2026 with the goal of evaluating the agreement between human scorers and our AI scoring model, using the responses obtained in the course of Study 1. 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.

Studies 1 and 2 each address the research objectives 1 and 2, as stated in the Introduction. Study 3 addresses the research objective 3.

3. Results

3.1. Psychometric properties of assessment and DIF analysis (Studies 1 and 2)

Using the DIF framework described above (with $p_{\text{threshold}}$=0.05), no DIF was found for any item in either Study.

The 2PL-IRT analysis found that the item difficulty levels cover a wide range without going to extremes (the range was approx. 0.2-0.6 in Study 1 and approx. 0.2-0.8 in Study 2), and the item discrimination was consistently very high (approx. 0.7-0.8 in Study 1 and approx. 0.7-0.85 in Study 2).

The internal consistency within each Power Skill scenario was found to be high: the lowest Cronbach's alpha was 0.92 in Study 1 and 0.95 in Study 2.

3.2. Agreement with human scoring (Study 3)

The main results were:[3]

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

[3] For calculating metrics that require 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 model scores, were rounded to the nearest such value.

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 similarly to human scorers. We note that the QWK between the mode and a human scorer is on average higher than among different human scorers.

3.3 Proficiency estimation

Having obtained the scores from Study 1, we pass them to the skill proficiency model, which estimates for each participant the posterior distribution of his or her latent proficiency on each skill in the framework. Our framework is a hierarchical structure (every Power Skill has multiple sub-skills in it). The proficiency estimates are produced for all levels of the hierarchy. By an interpretability convention, we multiply all proficiency values by 10, so that they are defined on the 0-10 range. Figures 3 and 4 show examples of such data.

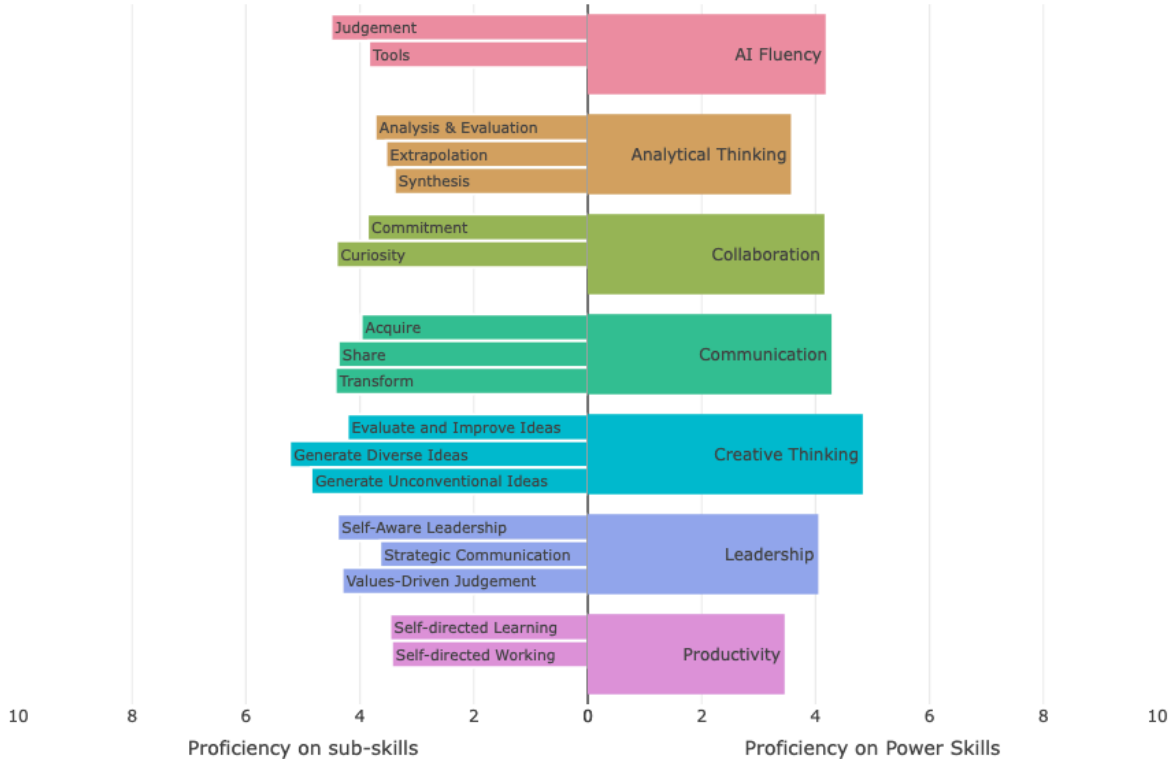

Figure 3. Reportable one-value proficiency estimates (means of posterior distributions) for Power Skills and their sub-skills for an individual.

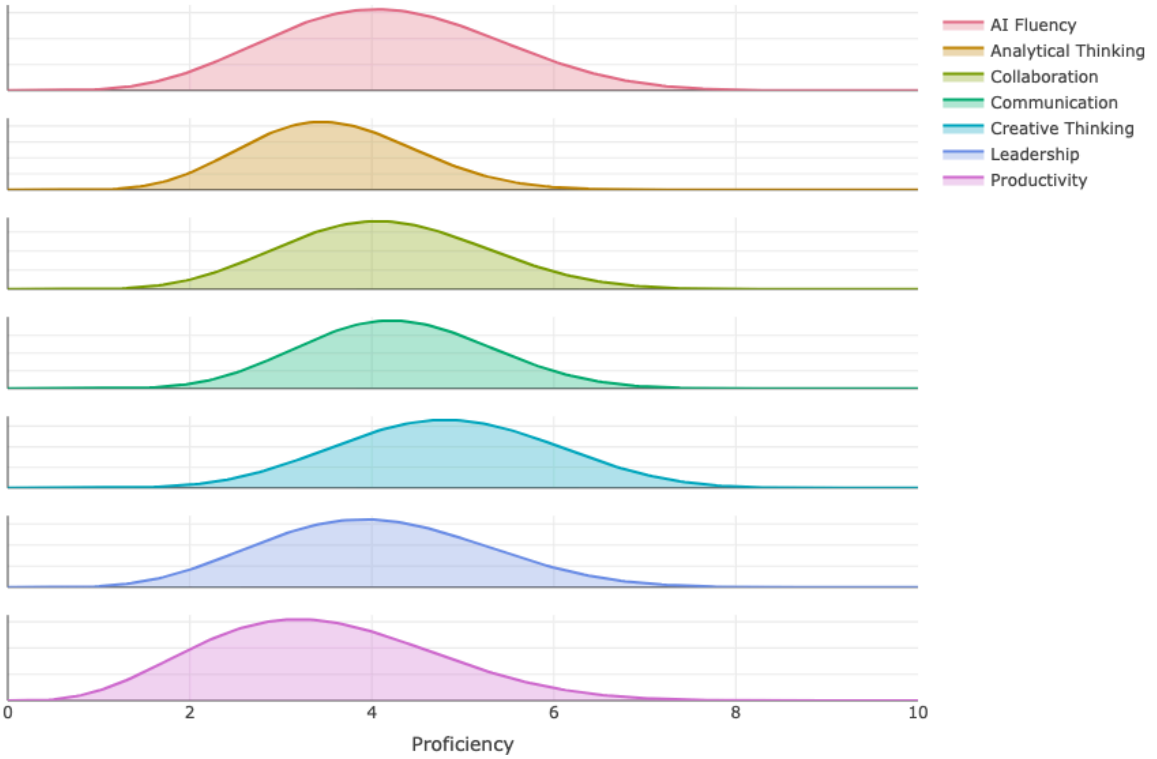

Figure 4. PDFs of proficiency posterior distributions of Power Skills for an individual.

We can demonstrate here one of many possible ways to use such data for practical purposes: calculating pairwise similarities among individuals in the proficiency space. This enables the identification of outlier individuals, as well as clustering individuals with the purpose of forming teams or groups in the development intervention activities. We use the Wasserstein W_2 metric to define the squared distance between two individuals:

$$D^2 = \sum_s \int_0^1 \left(Q_1^{(s)}(p) - Q_2^{(s)}(p)\right)^2 dp$$

where $Q_1^{(s)}$ and $Q_2^{(s)}$ are the quantile functions of the two individuals for a Power Skill s. The distance matrix D_{nm} obtained in this way (n and m enumerate individuals) can be used as the input for clustering, outlier-detecting, embedding and other models. Figure 5 shows a 2-dimensional configuration of a cohort of about 600 individuals, where visible distances approximate the distances defined above (a 2D embedding achieved by a standard multi-dimensional scaling). The individuals in this example have been split into 2 clusters by applying a hierarchical clustering that uses the Ward algorithm for inter-cluster distances, with the optimal number of clusters determined by the silhouette scores. Cross symbols denote individuals flagged as outliers by a local outlier factor model.

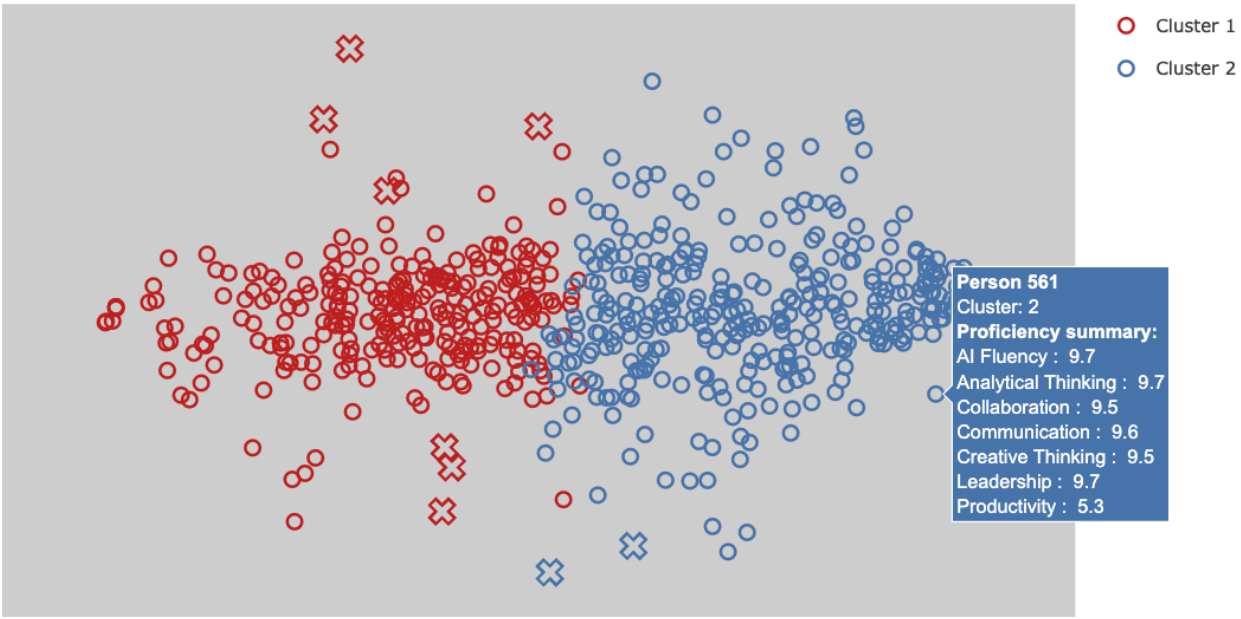

Figure 5. Visualization of a cohort of individuals that approximates distances among them. Crosses denote outliers identified by a local outlier factor model, which was applied to the full distance matrix and not to the approximate embedding, hence some of them may not look like outliers in this image.

4. Discussion

The purpose of this study was to evaluate the psychometric quality, fairness, and scoring validity of an AI-augmented assessment framework designed to measure human power skills through scenario-based, constructed-response tasks. Across three studies involving more than 1,000 participants, the results provide converging evidence that AI-enhanced assessment systems can support reliable, equitable, and scalable measurement of complex human competencies. Specifically, the findings demonstrated strong psychometric properties, no meaningful evidence of differential item functioning across major demographic groups, and substantial agreement between AI-generated scores and trained human raters. Together, these results support the feasibility of integrating advances in artificial intelligence and computational psychometrics into assessment systems intended to evaluate multidimensional human capabilities.

The findings contribute to a growing body of research suggesting that performance-based assessments can provide more authentic measures of human capability than traditional self-report instruments or selected-response tests. Contemporary workforce and leadership environments increasingly require individuals to demonstrate adaptive reasoning, communication, collaboration, creativity, and judgment in situations characterized by uncertainty and ambiguity. Consistent with evidence-centered design and learning engineering perspectives, the present framework captures observable manifestations of these capabilities through interactive scenarios rather than relying solely on self-perceptions or retrospective judgments. The strong item discrimination observed across both studies suggests that scenario-based tasks were effective in differentiating among participants with varying levels of underlying proficiency, while the broad distribution of item difficulty indicates that the assessments can capture performance across a wide range of ability levels.

The absence of meaningful differential item functioning across gender, race/ethnicity, and age groups represents an important finding in the context of responsible AI and assessment fairness.

Concerns regarding algorithmic bias have become central to discussions of AI-enabled talent management and workforce assessment. Although fairness can never be considered permanently established and requires continuous monitoring, the present results provide initial evidence that the assessment framework functions similarly across diverse demographic groups. These findings are particularly notable because the assessments relied on open-ended responses evaluated through AI-assisted scoring rather than conventional selected-response formats. The results therefore suggest that appropriately designed AI-augmented assessment systems can maintain fairness while extending measurement to more complex and authentic forms of performance.

The findings from Study 3 further strengthen the validity argument for AI-assisted scoring. Agreement between AI-generated scores and consensus human judgments exceeded average agreement among human raters themselves, indicating that the scoring models were capable of reproducing expert evaluation patterns with a high degree of consistency. This finding aligns with emerging research demonstrating that large language models can achieve substantial agreement with trained human scorers when evaluating constructed responses (e.g., Li et al., 2025). Importantly, the present framework does not position AI as a replacement for human judgment. Rather, it employs a human-in-the-loop architecture in which human expertise guides rubric development, calibration, quality assurance, and oversight. Such an approach reflects recommendations from the assessment literature that automated scoring systems should augment human expertise while preserving transparency, accountability, and psychometric rigor.

Beyond psychometric validation, the study highlights the potential value of Bayesian proficiency modeling for representing complex competency structures. Traditional assessment systems often reduce performance to single-point estimates that obscure uncertainty and relationships among skills. In contrast, the present framework generates posterior distributions of proficiency across a hierarchical skill ontology, enabling richer interpretations of individual and group performance. Such representations may support applications including personalized learning pathways, workforce development planning, leadership development, team composition, and organizational talent analytics. The ability to estimate multidimensional skill profiles and quantify uncertainty represents an important advancement in the measurement of human competencies that are inherently interconnected and context dependent.

An important implication of these findings concerns the measurement of communication and collaboration skills. Historically, these competencies have been difficult to assess at scale because evaluation often requires interpretation of nuanced discourse, interaction patterns, and contextualized social behaviors. Recent studies have demonstrated that large language models can reliably code collaborative communication according to established collaboration frameworks, achieving levels of agreement comparable to trained human coders. The present findings extend this literature by demonstrating that similar AI-assisted approaches can support scoring of broader power skills within realistic workplace and leadership contexts. Together, these results suggest that generative AI may enable scalable assessment of interpersonal and social-cognitive competencies that were previously constrained by the cost and complexity of human scoring. The results also raise important questions regarding how scoring reliability should be conceptualized in AI-augmented assessment systems (e.g., Jung, Bezirhan, & von Davier, 2025). Traditional approaches rely heavily on inter-rater agreement and double scoring. Emerging work in computational psychometrics suggests that semantic consistency among similar responses may provide an additional lens for evaluating scoring quality. From this perspective, reliability is not solely a property of agreement between raters but also reflects whether semantically similar responses receive similar scores. Future work could integrate semantic similarity auditing methods with human–AI agreement analyses to provide more comprehensive evidence regarding scoring consistency, particularly for complex competencies such as communication, collaboration, leadership, and creative thinking.

The study also has broader implications for assessment in the age of generative AI. As AI systems become increasingly capable of producing sophisticated written responses, traditional approaches to measuring human capability face growing validity challenges. Assessment systems must move beyond evaluating static outputs toward capturing evidence of reasoning processes, decision making, adaptation, and judgment within authentic contexts. The present framework represents one possible response to this challenge by leveraging AI not only as a scoring mechanism but also as an interactive assessment partner capable of eliciting evidence of human thinking and behavior through dynamic, multi-turn scenarios. This approach may help preserve the relevance of competency assessment in environments where AI assistance is increasingly ubiquitous.

Several limitations should be considered when interpreting the findings. First, the participant samples were recruited through a crowdsourcing platform and, although demographically diverse, may not fully represent organizational populations, executive leaders, or international workforces. Additional validation studies in operational settings will be necessary to establish generalizability across industries, cultures, and organizational contexts. Second, the current studies focused primarily on internal psychometric properties, fairness indicators, and scoring agreement. Future research should examine predictive validity, including the extent to which assessed power skills relate to workplace performance, leadership effectiveness, learning outcomes, innovation, and career progression. Third, although no evidence of differential item functioning was detected, fairness is an ongoing process rather than a static outcome. Continuous monitoring, periodic recalibration, and independent auditing will remain important as assessment content, AI models, and user populations evolve. An additional limitation concerns the scope of the fairness evidence presented here. Although differential item functioning analyses found no meaningful evidence of demographic bias, recent work on AI-assisted scoring suggests that subgroup consistency should be evaluated across multiple dimensions, including scoring agreement, reliability, and predictive relationships with external criteria. Future research should therefore examine these additional forms of subgroup consistency using larger and more diverse samples, particularly when assessing communication, collaboration, and leadership behaviors that may be expressed through culturally variable interaction styles. Finally, the present investigation focused on English-language assessments and interactions. Additional work is needed to examine cross-linguistic validity and cultural fairness in multilingual environments.

Future research should explore several promising directions. Longitudinal studies could investigate the sensitivity of AI-augmented assessments to skill growth over time and their utility for measuring the impact of educational and workforce development interventions. Additional research may examine adaptive assessment mechanisms that dynamically adjust scenario complexity based on participant performance. Further work is also needed to investigate how human–AI collaboration itself should be incorporated into competency models, particularly as AI fluency becomes an increasingly important component of professional effectiveness. Finally, future studies should evaluate how probabilistic skill profiles can be integrated into talent development, workforce planning, and organizational decision-making systems while maintaining fairness, transparency, and ethical governance.

Overall, the findings provide evidence that AI-augmented assessment systems can achieve psychometric standards comparable to traditional measurement approaches while extending assessment to complex, authentic, and multidimensional human capabilities. By integrating advances in generative AI, computational psychometrics, and evidence-centered assessment design, the

framework offers a scalable approach to measuring the competencies that remain essential in an increasingly AI-enabled world.

The present work also contributes to the emerging field of AI-native assessment, in which artificial intelligence is not merely used to automate existing testing processes but is integrated into the design, delivery, scoring, and interpretation of assessment itself. Traditional digital assessments often replicate paper-based paradigms in electronic formats, whereas AI-native assessments leverage interactive, adaptive, and conversational environments to elicit richer evidence of human reasoning, communication, collaboration, and judgment. Within this paradigm, AI serves simultaneously as an assessment facilitator, evidence collector, scoring engine, and psychometric component of the measurement system. The framework presented in this study demonstrates how advances in generative AI can be combined with established principles of psychometrics, evidence-centered design, and human oversight to create assessment systems that are both scalable and scientifically defensible. As AI becomes increasingly embedded in educational and workplace contexts, AI-native assessment may provide a foundation for measuring the uniquely human capabilities that remain critical for effective performance in partnership with intelligent technologies.

5. Conclusions

This research evaluated an agentic AI assessment framework designed to measure human power skills through scenario-based, constructed-response interactions and human-in-the-loop AI scoring. Across three empirical studies involving more than 1,000 participants, the framework demonstrated strong psychometric performance, including high item discrimination, broad coverage of difficulty levels, strong internal consistency, and no meaningful evidence of differential item functioning across demographic groups. In addition, AI-generated scores showed substantial agreement with trained human raters and achieved higher consistency with consensus judgments than the average agreement observed among human scorers.

These findings provide evidence that agentic AI assessment systems can support reliable, fair, and scalable measurement of complex human competencies. The results suggest that advances in generative AI and computational psychometrics can be leveraged not only to automate assessment processes but also to strengthen the measurement of capabilities that are difficult to evaluate using traditional testing methods. By combining interactive scenario-based assessment, AI-assisted scoring, and Bayesian proficiency estimation, the proposed framework extends competency measurement beyond static tests and self-report instruments toward more authentic representations of human performance.

As organizations, educational institutions, and workforce systems increasingly seek to understand and develop human capabilities that complement artificial intelligence, robust methods for measuring these competencies become increasingly important. The present findings suggest that AI can play a constructive role in this process when embedded within psychometrically grounded, transparent, and human-supervised assessment systems. More broadly, the study demonstrates that scalable measurement of human power skills is both technically feasible and scientifically defensible, providing a foundation for future research and practical applications in talent development, workforce readiness, leadership assessment, and lifelong learning.

References

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces collective diversity. *Science Advances, 10*(5).

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S. M., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice, 43*(2), 16–29.

Jung, J. Y., Bezirhan, U., & von Davier, M. (2025). Reconceptualizing scoring reliability through linguistic similarity. *Educational and Psychological Measurement*.

Figalli, A., & Glaudo, F. (2021). *An invitation to optimal transport, Wasserstein distances, and gradient flows* (Vol. 1). EMS press.

Halpin, P. F. (2024). Differential item functioning via robust scaling. *Psychometrika*, 89(3), 707–736.

Kleinberg, J., & Raghavan, M. (2021). How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation, 9*(4), 1–23.

Li, H., Chen, C. H., Fan, K., Young-Johnson, C., Lim, S., & Feng, Y. (2025). *Agreement between large language models and human raters in essay scoring: A research synthesis*. arXiv preprint.

National Academies of Sciences, Engineering, and Medicine. (2024). *Foundational research on artificial intelligence in education and workforce development*. National Academies Press.

Rosen, Y. (2015a). Measuring collaborative problem solving through technology-based assessments. *Educational Measurement: Issues and Practice, 34*(1), 16–28.

Rosen, Y. (2015b). 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. arXiv preprint.

Rosen, Y., & Rushkin, I. (2026b). *Proposed Framework for Agentic AI Assessment of Creative Thinking for the U.S. Air Force*. Paper 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.

Rosen, Y., & Rushkin, I. (2026c). *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). T*he 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.

Teresi, J. A., Jones, R. N., Kleinman, M., Ocepek-Welikson, K., & Morales, L. S. (2021). Modern methods for detecting differential item functioning: Applications and future directions. *Psychological Test and Assessment Modeling*, 63(1), 1–32.

von Davier, A. A. (2024). How will AI change adaptive testing? In D. Yan, C. Lewis, & A. A. von Davier (Eds.), *Research for practical issues and solutions in computerized multistage testing* (1st ed., pp. 363–384). Routledge.

von Davier, A. A., Mislevy, R. J., & Hao, J. (2023). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python*. Springer Nature.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.