Workera

Technical Overview:
Workera's AI-Native Elo Assessment Platform

Authored by **Taylor Sullivan,** PhD, Head of Assessments @ Workera

V 1.0

Executive Summary

This paper describes the psychometric framework underpinning Workera's Elo platform, an AI-native system for skill assessment and development. Drawing on principles of evidence-centered design and modern learning science, we examine the platform's approach to scoring, item generation, adaptivity, validity, fairness, and reliability.

In contrast to traditional psychometric paradigms such as computer adaptive testing (CAT), a legacy method that adapts item difficulty based on response accuracy but is limited in its ability to capture multidimensional skills, context sensitivity, or real-world job alignment, Elo integrates artificial intelligence across the assessment lifecycle, enabling flexible, context-rich measurement aligned with contemporary workforce needs.

Authored by
Taylor Sullivan
PhD, Head of Assessments @ Workera

Workera

Overview by Section

In the following sections, we outline the foundational principles, methodologies, and innovations behind the Elo assessment platform. Each section highlights a core element of Elo's psychometric framework—demonstrating how the platform delivers rigorous, adaptive, and future-ready skill measurement. The report is organized as follows:

01.
Introduction

Elo was developed in response to the growing need for more flexible, adaptive, and job-relevant skill assessment systems. Traditional models, including those based on CAT, often rely on fixed item banks, static scoring models, and pre-calibrated psychometric parameters that are slow to evolve with industry demands. Elo introduces a fundamentally different approach, integrating artificial intelligence with human expertise to build an assessment ecosystem designed for the realities of modern workforce development.

This paper provides an overview of the psychometric framework supporting Elo, with specific focus on model establishment and scoring, item development and calibration, adaptivity, validity, fairness, reliability, and personalized learning.

02.
Item Development and Calibration

With Elo, Workera takes a fundamentally different approach to item development and adaptation, leveraging AI-driven content generation, human refinement, and continuous model iteration to ensure assessments remain relevant, scalable, and aligned with evolving skill demands. Unlike traditional assessments that rely on pre-calibrated item banks and static parameter estimation, Elo allows for dynamic item creation and real-time difficulty adjustments based on user interactions.

Important to note is that all items are mapped to structured competency frameworks and undergo validation through both real-user pilots and synthetic user simulations. Each item is explicitly linked to a specific skill within Elo's modular skill taxonomy, ensuring alignment with well-defined constructs. During development, items are reviewed by humans and AI agents for both technical accuracy and construct relevance, and then piloted in controlled settings to evaluate quality prior to operational use.

Content Development Workflow

- **AI-Assisted Item Generation:** Elo generates assessment content mapped to structured competency models, ensuring each item aligns with a well-defined skill. This process allows for rapid iteration, scalability, and the ability to quickly update assessments as industry standards evolve.
- **Human Refinement & Expert Review:** Subject matter experts (SMEs) review and refine AI-generated items to ensure accuracy, fairness, and job relevance. SMEs help structure prompts, verify alignment with real-world skill applications, and enhance question clarity.
- **Expanded Item Formats:** Elo's items go beyond multiple-choice questions (MCQs), incorporating short-answer responses, coding tasks, interactive whiteboarding, case-based reasoning, and other applied skill demonstrations. For example, a leadership task might involve whiteboarding a response strategy during a simulated crisis, while a data science task could require prioritizing model evaluation metrics in a given business context. This approach better reflects how skills are used in practice rather than just testing for rote knowledge.
- **Automated Quality Control:** Our embedded quality control copilot helps ensure assessments remain accurate, reliable, and fair by continuously monitoring item performance. This copilot is powered by AI, continuously reviewing item performance metrics and detecting questions that are too easy or difficult, response time anomalies, and inconsistent answer patterns, allowing assessment developers to analyze data and explore root causes. Developers can investigate whether issues stem from poor wording, ambiguous concepts, or unintended biases and refine items accordingly.

Calibration Without Traditional Parameter Estimation

Unlike traditional assessments that rely on traditional parameter estimation, Elo assessments are not constrained by pre-calibrated item banks or static IRT models requiring fixed statistical parameters like difficulty and discrimination. Instead, Elo uses statistical modeling of historical response patterns, expert-validated rubric scoring, and predictive scoring simulations via calibrated synthetic user profiles. These tools maintain psychometric rigor while offering greater flexibility.

- **Data-Informed Calibration:** Elo leverages AI-powered scoring and historical response data to guide item development and maintain consistency in measurement. While not using traditional cut scores, past performance trends help shape difficulty expectations for new items.
- **Prioritizes Skill Profiling Over Fixed Cut Scores:** Instead of relying on IRT-based ability estimates, Elo scoring is skill-based, mapping responses to proficiency classifications using AI inference models.
- **Continuously Refines Models:** Unlike traditional assessments that require batch recalibration of item parameters, Workera's models underpinning Elo continuously adapt based on synthetic user testing, real-world response data, and iterative AI learning, keeping assessments aligned with evolving skill demands.

Workera

03. Adaptive Testing Methodology

Elo's **AI-driven adaptive assessments** offer a more flexible and nuanced approach to skill measurement. While traditional CAT systems adapt based on a user's estimated ability level, Elo employs a multi-dimensional, session-aware adaptivity framework that integrates contextual and behavioral signals. Instead of following a predefined algorithm that selects isolated items based on estimated ability, Elo adapts holistically by considering response patterns and learning signals within a session. This approach creates richer interactions, deeper skill insights, and a more engaging test-taker experience while aligning with modern learning science and skill development models.

Multi-Turn Item Structures

Traditional CAT adjusts item difficulty sequentially but treats each question as independent.
Elo structures **multi-turn item interactions**, enabling:

- **Context Preservation:** Multi-turn items allow related questions to share context, reducing cognitive switching and enabling deeper skill exploration.
- **Reduced Cognitive Load:** Multi-turn items share a common context across related questions, allowing test-takers to stay engaged without constantly shifting mental frameworks between unrelated tasks.

Example: In a systems design assessment, a user might first sketch out a high-level architecture, followed by follow-up questions asking them to justify trade-offs or adjust based on changing requirements.

Response-Aware Adaptivity

- **Richer Skill Signals:** Instead of assessing proficiency with isolated responses, Elo enables deeper exploration of reasoning, problem-solving strategies, and applied understanding through follow-up questions and adaptive task progression.
  Importantly, factors like response time, grammar, or phrasing style are not used to penalize scores unless they are construct-relevant (e.g., grammar in a writing skill test). Elo separates stylistic expression from skill demonstration and has safeguards to avoid unintended scoring bias.
- **Dynamic Adaptation to User Responses:**
  - High-proficiency users receive progressively complex challenges that test deeper problem-solving and application.
  - Users demonstrating uncertainty (e.g., multiple "I don't know" responses) may be provided scaffolding, clarification, or an early exit option to minimize frustration and optimize cognitive effort.
  - Future iterations may explore adjusting the question framing, modality (text, code, interactive tasks), or feedback approach based on response patterns.

Beyond One-Dimensional Ability Estimation

Traditional CAT optimizes for efficiency in estimating a single latent ability trait, typically maximizing information at a given theta level. Elo, by contrast, is **not confined to a single ability estimation model.** Instead, it accounts for skill-specific response patterns, contextual performance signals, and task-based reasoning to provide a more complete picture of user proficiency.

A More Natural and Engaging Assessment Experience

Traditional CAT follows a strict item-by-item selection process that can sometimes feel mechanical or test-like rather than reflecting how people demonstrate skills in real-world problem-solving scenarios. Elo's approach preserves measurement rigor while enabling more fluid and meaningful assessment experiences that better align with applied skill demonstration.

04.
Scoring Architecture and Skill Interpretation

Elo employs a novel, AI-native scoring architecture designed to move beyond traditional binary item scoring. Rather than assigning fixed point values to right or wrong answers, Elo constructs multi-dimensional skill profiles that integrate both directly observed and inferred evidence of competence. This framework supports robust, interpretable, and adaptive measurement of user capabilities grounded in educational and psychological best practices.

Scoring Pipeline: AI + Human Collaboration

Elo uses an AI-driven, expert-informed scoring approach. Large language models (LLMs) score responses using pre-defined rubrics, while human experts shape scoring frameworks, refine prompts, and validate scoring on sample responses. During development, AI scoring is benchmarked against human judgment and subjected to regular internal review, particularly for open-ended domains.

This integrated model ensures scalability without compromising measurement integrity. Every rubric is:

- Mapped to **granular skills within structured competency models**
- Anchored by **benchmark responses** to ensure alignment and interpretability

AI assigns each skill as either "Strong" (meeting expectations) or "Gap" (falling short). Plans for partial credit scoring (e.g., multi-step reasoning or code tasks) are in development.

Skills Classification Framework

At the core of Elo's scoring methodology is a four-part skill classification system that distinguishes between directly assessed and inferred competencies:

Classification	Description	Evidence Type	Stakeholder Utility
Verified Skill	Directly assessed and meets mastery criteria defined by SMEs	AI-scored rubric-based	High-confidence confirmation of capability
Potential Skill	Inferred based on strong performance in adjacent or prerequisite areas	Inferred via Deep Knowledge Tracing	Broader profile coverage with estimated accuracy
Verified Gap	Directly assessed and falls below mastery thresholds	AI-scored rubric-based	High-confidence indicator of skill development need
Potential Gap	Inferred as likely underdeveloped due to weak signal in related competencies	Inferred via DKT	Priority area for learning or follow-up evaluation

This classification system enables both formative and summative interpretations, supporting diagnostic feedback and precision workforce capability mapping.

Domain Proficiency Estimation

Elo aggregates skill-level classifications into domain proficiency scores, reported on a standardized 1–300 scale. This scale supports alignment with organizational proficiency frameworks and simplifies longitudinal tracking.

Score Range	Proficiency Band	Interpretation
1-100	Beginning	Foundational awareness
101-200	Developing	Emerging, partial proficiency
201-300	Accomplished	Consistent, advanced capability

Planned enhancements include the introduction of error bands or confidence intervals around proficiency estimates to reflect inherent measurement uncertainty, especially for shorter or highly adaptive assessments. These error metrics will support more precise interpretation of scores at both the individual and cohort levels.

In parallel, Workera is investing in psychometric monitoring strategies to evaluate scoring consistency across different user segments, leveraging both real user data and synthetic personas. This monitoring helps detect drift or potential scoring disparities across industries, geographies, or organizational populations—supporting fairness, reliability, and interpretability at scale.

Skill Inference: Deep Knowledge Tracing (DKT)

Elo integrates a proprietary Deep Knowledge Tracing model to infer skill mastery from observed behaviors. DKT creates a latent skill graph based on response patterns, skill co-occurrence, and contextual embeddings, enabling estimation of likely proficiency for untested skills. DKT functions like a dynamic mental map—connecting what a learner demonstrates with what they likely know, even if it hasn't been directly tested—similar to how a mentor or tutor infers mastery from indirect evidence.

> **Example:** A user who performs well on Transformer-based tasks may be inferred to understand foundational neural network concepts, even if those items weren't directly tested.

This allows Elo to reduce assessment length while maintaining broad profile coverage—a key advantage for scalable diagnostics.

Interpretability and Auditability

Each skill classification in Elo is:

- **Traceable:** Backed by response-level data or inference logic
- **Auditable:** Scoring and inference methods are documented to support review and QA
- **Actionable:** Mapped to feedback and learning pathways

Elo ensures measurement consistency through data-informed calibration, expert-guided scoring development, and real-world performance monitoring. Rather than relying on fixed IRT parameters, scoring models evolve based on live data, calibrated synthetic user validation, and domain expert input.

This dynamic scoring architecture supports flexible, rigorous, and future-ready skill measurement—anchored in defensible methodology and optimized for practical application.

05.
Validity and Intended Score Use

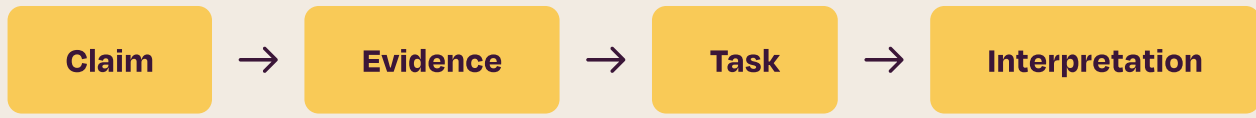
Our validation approach is designed to support **skill assessments where the primary goal is to signal proficiency in a specific domain.** In this context, we are not making predictive claims about job performance, nor are our assessments intended to serve as selection instruments. Instead, the assessment is optimized to provide **reliable, interpretable signals of skill strength** within a dynamic upskilling or development ecosystem. As such, Elo is most appropriate for learning and development contexts, skills diagnostics, and role-readiness mapping where the goal is to inform rather than to select.

This intended use shapes both how we build our assessments and the types of validity evidence we prioritize. We focus on:

Content validity: ensuring alignment between what is measured and real-world skills
Construct validity: ensuring the interpretation of scores accurately reflects skill proficiency
Procedural validity: ensuring tasks elicit the intended skills through realistic demonstrations
Continuous validation: ensuring our assessments evolve with workforce needs

While we do not make predictive claims, we recognize that clients may apply Elo scores in varied contexts. In such cases, organizations are encouraged to **conduct local validation studies** to ensure alignment with their internal frameworks, regulatory obligations, or job-role expectations.

A Use-Aligned, Evidence-Centered Approach



Claim

The learner is proficient in a defined skill domain—e.g., Machine Learning Foundations—and is ready to apply core concepts in a professional context (e.g., building models, evaluating performance, interpreting results).

Evidence

- **Direct evidence:** High-quality responses to tasks measuring core ML skills (e.g., supervised vs. unsupervised learning, confusion matrices).
- **Inferred evidence:** Workera's Deep Knowledge Tracing (DKT) model predicts proficiency in adjacent skills (e.g., ROC analysis, regularization) based on observed performance and latent skill relationships.

Elo assessments are grounded in **evidence-centered design (ECD)** (Mislevy et al., 2003, 2018), a framework that ensures every assessment is built around a clearly defined claim and structured to gather appropriate evidence through meaningful tasks. This design philosophy promotes interpretability, validity, and defensibility.

Here's how this framework is applied in practice:

Example: A learner who accurately explains overfitting and selects appropriate cross-validation methods may be inferred to understand bias-variance tradeoffs—even without explicit testing—based on skill graph co-occurrence patterns.

Task

Learners engage in applied, scenario-based tasks such as:

- Explaining overfitting and mitigation strategies
- Selecting and justifying model evaluation metrics based on real-world business constraints
- Interpreting model outputs and recommending next steps in a machine learning workflow

These tasks are designed by SMEs in conjunction with authoring LLMs, mapped to granular competencies, and scored using structured rubrics supported by AI and human validation.

Interpretation:

- A domain score of ~250 (out of 300), with 87% or more of skills classified as "Verified", would indicate the learner has demonstrated sufficient breadth and depth of skill to contribute meaningfully in professional ML environments.
- If "Verified Gaps" appear in areas like data preprocessing or hyperparameter tuning, the system flags these as targeted upskilling priorities.
- If "Potential Gaps" are detected in advanced topics (e.g., dimensionality reduction), the interpretation may suggest the learner is competent at the foundational level but not yet ready for specialized or production-grade modeling work.

This interpretation supports personalized learning, role-readiness diagnostics, and career mobility planning—without making unjustified predictive claims about future job performance.

Why Elo's Validity Model Differs from Legacy Approaches

Unlike traditional approaches to validation, which assume both the predictor (assessment scores) and criterion (job performance) remain stable over time, Elo's AI-native approach acknowledges that **both skills and job expectations evolve dynamically.**

Elo's **strengths in validity** stem from its ability to align directly with evolving workforce skills rather than relying on static validation studies tied to outdated job requirements.

Our approach:

- **Is Built for Dynamic Skills Markets:** Traditional assessments validate against historical performance benchmarks that often lag behind industry change. Elo's AI-generated competency models are designed to evolve based on expert input and industry feedback, ensuring that the construct measured remains relevant. Elo recognizes that skill proficiency is not just about hitting a single benchmark but about continuous learning and adaptation—meaning that assessments must evolve alongside the skills they measure.
- **Prioritizes Applied Skill Demonstration Over Abstract Measurement:** Many traditional assessments focus on generalized cognitive ability or domain knowledge rather than task-based skill application. Elo assessments are designed to capture how individuals interact with real-world scenarios, making our construct validity strongly aligned with job performance expectations.
- **Combines AI-Generated and Expert-Refined Validation:**
  - **Construct Validity:** AI-generated competency models are mapped to measurable, job-relevant skills and iterated upon by SMEs to refine constructs and ensure assessment accuracy.
  - **Expert-Driven Content Validation:** SMEs play an active role in reviewing, refining, and validating assessments, ensuring fairness and industry relevance. We also collect user feedback post-assessment to monitor alignment with real-world expectations.
  - **Continuous vs. One-Time Validation:** With Elo, Workera treats validation as **a living process** rather than a fixed historical benchmark. In today's workforce, skill requirements shift rapidly, and the definition of job success evolves. A study conducted today may not be predictive in just a few years—yet traditional models treat validation as a one-time event. Rather than assuming that one fixed validation study will establish enduring accuracy, we treat **validation as a continuous process**, ensuring that assessments reflect real-world job expectations as these expectations evolve.

By integrating **AI/ML, psychometric best practices, and real-world task modeling,** Elo provides validity evidence that is purpose-aligned, transparent, and adaptable.

06.
Fairness and Integrity

Workera is committed to **ensuring fairness in assessment outcomes** while maintaining a **privacy-first approach** that aligns with evolving regulatory and political considerations. Our AI-driven fairness mechanisms are designed to identify and mitigate bias without requiring demographic data collection, ensuring that our assessments remain inclusive, responsible, and adaptable. Our approach includes:

AI-Driven Fairness Detection & Human Oversight

- **AI-Based Fairness Detection:** AI models undergo automated bias and sensitivity checks, leveraging linguistic, semantic, and contextual analysis to flag potentially problematic content. These models are trained to detect cultural, gendered, or exclusionary language patterns that could introduce unintended bias.
- **Expert Oversight & Iteration:** Humans review AI-flagged items, refining prompts, adjusting context, and ensuring assessments adhere to fairness principles without sacrificing measurement precision.

Procedural Integrity

To ensure that Elo assessments yield trustworthy, interpretable results, Workera maintains safeguards focused on **session integrity and secure delivery within our platform environment.**

- **Assessment Access Controls:** Assessments are distributed through authenticated enterprise platforms and gated interfaces. Users cannot access content outside of a sanctioned environment, reducing unauthorized sharing or misuse.
- **Scoring Environment Security:** All scoring is performed within Workera's secure AI infrastructure—meaning neither item content nor scoring logic is exposed during or after the session.
- **Session Tokenization and User Authentication:** Assessment sessions are uniquely tied to authenticated user accounts via session tokens, ensuring that only intended users can initiate, resume, or submit a given assessment. This prevents impersonation, session hijacking, or unauthorized access through shared links or stale credentials.
- **Frontend Shielding of Assessment Logic:** Test logic, including scoring conditions, adaptive rules, and branching structures, are executed server-side within Workera's secure environment. No sensitive assessment logic is exposed on the client side, reducing the risk of reverse engineering or exploitation.

Privacy-First Approach & Regulatory Considerations

- **Navigating an Evolving Political Landscape:** Given the current political landscape, where federal grants and contractors face increasing scrutiny over demographic data collection, Workera has strategically chosen not to collect such data for the time being to mitigate the risk of potential funding restrictions and regulatory challenges.
- **Empowering Customers to Conduct Fairness Audits:** While Workera does not collect demographic data internally, we enable customers to conduct their own fairness analyses by providing score data exports. This allows organizations to analyze assessment outcomes through their own demographic lenses, ensuring compliance with internal fairness objectives and external regulatory requirements. In turn, this better supports flexibility for different industries and geographies to apply fairness evaluations aligned with their unique requirements. Workera provides clients with user-level performance data on each skill and assessment. Clients can pair these results with internal demographic datasets to conduct DEI audits. This privacy-respecting approach allows fairness insights without centralized demographic storage.

07.
Reliability

Elo ensures **consistent and precise skill measurement** through AI-driven monitoring, synthetic user testing, and continuous calibration. Rather than relying on fixed-item calibrations or traditional test-retest methods, our approach prioritizes ongoing model refinement to maintain reliability while allowing for adaptability. We leverage:

- **Synthetic User Testing:** AI-generated personas simulate assessment responses to evaluate item consistency, scoring stability, and response patterns before assessments are deployed to real users.
- **Continuous Model Refinement:** AI tracks scoring consistency in real time, detecting anomalies and refining scoring models to prevent drift or misalignment over time. Experts intervene as needed to audit and adjust model parameters.

Forthcoming Reliability Metrics

As of today, Elo does not report traditional reliability coefficients (e.g., Cronbach's alpha or test-retest reliability). This is not due to a lack of psychometric rigor, but rather because **conventional reliability metrics are difficult to compute or interpret meaningfully in the context of Elo's AI-native, adaptive, and multi-format assessment model.**

Several design factors contribute to this challenge:

- **Adaptive Item Selection:** Elo will be able to dynamically adjust item sequences based on user responses, meaning users may see different sets of tasks. This violates the fixed-form assumptions required for internal consistency metrics.
- **Mixed Item Formats:** Elo combines diverse item types within a single assessment—such as coding exercises, open-ended responses, whiteboarding simulations, and scenario-based reasoning tasks. These item types are not interchangeable in content or cognitive demand, making internal consistency measures like Cronbach's alpha inappropriate.
- **Latent Inference Models:** Elo uses Deep Knowledge Tracing to infer mastery of untested skills, further complicating the use of classical reliability metrics that assume directly observed responses for all measured traits.

Despite these constraints, **Workera is actively developing modern reliability evidence** suited for AI-native assessments. For example, we are exploring generalizability theory analyses, semantic and conceptual consistency metrics, and SME-based scoring audits.

As these approaches mature, Workera will publish **construct-aligned reliability indicators** that are transparent, interpretable, and relevant to the specific nature of each assessment.

Workera

08.
Personalized Learning Integration

A key differentiator of Elo's approach is that assessments are not standalone evaluations but are **directly integrated into adaptive learning experiences.** Unlike traditional assessments that focus solely on measuring ability, Elo bridges the gap between assessment and development, ensuring that users receive targeted, actionable learning recommendations that support continuous skill growth.

How Elo Translates Assessment Into Personalized Learning

- **AI-Driven Skill Gap Analysis:** Assessment results are automatically analyzed to identify areas of strength and opportunity, generating personalized upskilling recommendations tailored to each individual's proficiency profile.
- **Adaptive Recommendation Engine:** Rather than prescribing static learning modules, Elo dynamically adjusts next-step learning recommendations based on assessment performance, prioritizing skills with the highest impact on career progression.
- **Seamless Integration with Learning Platforms:** Elo's recommendation system is designed to integrate with enterprise L&D platforms, personalized coaching programs, and third-party learning ecosystems, ensuring a cohesive, scalable learning experience. Elo outputs can integrate with platforms such as Degreed, Workday, or internal LMSs via export APIs or direct connectors, enabling alignment with existing coaching workflows and learning content libraries.
- **Planned Enhancement: Multi-Modal Learning Recommendations:** Based on assessment results, users receive diverse, high-quality learning interventions, including curated courses, interactive challenges, hands-on projects, and applied exercises aligned with their skill needs.

Conclusion

Elo represents a departure from conventional psychometric models, integrating artificial intelligence with human expertise to support dynamic, scalable, and context-aware skill measurement. Unlike traditional assessments that rely on fixed structures, static validity studies, and singular ability metrics, Elo is designed to evolve alongside the skills it measures.

By integrating AI-driven skill inference, real-time adaptation, expert validation, fairness monitoring, and targeted learning pathways, Elo establishes a foundation for skill assessment that aligns with the complexity of modern work. Our approach is designed to maximize engagement, optimize learning outcomes, and provide organizations with deeper insights into workforce capabilities—all while maintaining measurement rigor, fairness, and reliability without the constraints of legacy psychometric models.

Version	Release Date	Comments
1.0	**April 2025**	**Initial release by Taylor Sullivan, PhD**