International Conference on Education, Psychology and Humanities
Hosted Online from Moscow, Russia
Date: 28th February, 2026
Website: https://econferencia.com

AI-ASSISTED SCORING OF FOREIGN LANGUAGE PROFICIENCY EXAMS: PROBLEMS AND SOLUTIONS

Kamalova Shakhlo Nugmanovna
Senior lecturer at Sarbon University in Tashkent

Abstract

The use of artificial intelligence (AI) to score foreign language proficiency exams is expanding due to the need for scalability, faster turnaround, and cost efficiency—especially in large-scale, high-stakes contexts. AI-assisted scoring commonly includes automated marking of selected-response items, automated essay scoring (AES) for writing, and speech technologies (automatic speech recognition and pronunciation/prosody models) for speaking assessments. Despite operational benefits, AI scoring introduces critical challenges: construct underrepresentation, bias and fairness risks across accents and demographic groups, limited explainability, vulnerability to gaming, domain shift across prompts and test forms, and governance issues related to data privacy and accountability. This article synthesizes key problems associated with AI-based scoring and proposes a solutions framework centered on construct validity, human-in-the-loop moderation, rigorous psychometric calibration, continuous bias auditing, robust security controls, and transparent candidate-facing policies (including appeal mechanisms). The paper argues that AI can be used responsibly in language assessment only when it is embedded within a defensible assessment design that prioritizes validity, reliability, and equity.

Keywords: AI scoring; automated essay scoring; speech assessment; language testing; validity; fairness; bias auditing; human-in-the-loop; CEFR; high-stakes exams.

International Conference on Education, Psychology and Humanities
Hosted Online from Moscow, Russia
Date: 28th February, 2026
Website: https://econferencia.com

Introduction

Foreign language proficiency exams play a gatekeeping role in education, employment, and migration. Their results often carry high stakes for test takers, which makes scoring practices subject to strict requirements for reliability, validity, fairness, transparency, and accountability. Traditional scoring approaches—human raters for writing and speaking, and machine scoring for selected-response items—have been increasingly augmented by AI methods. Recent advances in natural language processing (NLP) and speech processing have made it feasible to automate or partially automate scoring of constructed responses at scale.

AI-assisted scoring can reduce costs, shorten reporting timelines, and improve consistency when compared with purely human scoring systems that may suffer from rater drift, fatigue, and variability. However, language ability is a complex construct: it includes accuracy, fluency, coherence, pragmatic appropriateness, and task achievement. When AI models are used to predict scores, they may rely on surface features (length, lexical frequency, acoustic clarity) that correlate with proficiency but do not fully represent communicative competence. This creates a risk that automated scores reflect "what the model can measure" rather than "what the test intends to measure."

Additionally, AI scoring systems are sensitive to accent variation, background noise, topic familiarity, and the demographic distribution of training data. These factors can produce systematic score differences unrelated to proficiency, undermining fairness. Because AI decisions can be opaque, candidates and institutions may also struggle to understand or contest outcomes.

This paper examines AI-assisted scoring for foreign language proficiency exams, focusing on major problem categories and actionable solutions. The central claim is that responsible deployment requires an integrated approach: assessment

design and psychometrics must govern model development—not the other way around.

Methodology

This article uses a conceptual-analytical methodology grounded in established principles of language assessment and educational measurement. The analysis is organized as follows:

1. **System decomposition:** AI scoring is separated into (a) selected-response scoring, (b) writing scoring (AES), and (c) speaking scoring (ASR + scoring models).
2. **Validity-centered evaluation:** risks are mapped to threats to validity (construct representation, irrelevant variance, and generalization).
3. **Fairness and governance lens:** issues are classified into technical (bias, robustness), operational (moderation, appeals), and legal/ethical (privacy, accountability) categories.
4. **Solutions framework:** mitigations are proposed at three layers—assessment design, model development, and operational deployment—emphasizing measurable controls (audits, calibration, monitoring).

No new empirical dataset is introduced; instead, the paper synthesizes established findings in the literature and translates them into an implementable checklist for exam providers.

Research on automated scoring predates modern generative AI. Automated essay scoring has long used features such as grammar, lexical diversity, cohesion, and discourse structure, with later approaches shifting toward machine learning and neural representations. In parallel, speaking assessment has leveraged ASR outputs (word error rate, disfluency, speaking rate) and acoustic/prosodic features. While these systems can approximate human ratings under controlled

conditions, they face known limitations: they may reward formulaic writing, penalize non-native accents, and degrade when prompts, topics, or populations differ from training data.

In language testing theory, the dominant concern is **validity**: whether score interpretations are justified for their intended uses. Reliability (score consistency), fairness (absence of systematic disadvantage), and comparability (across forms and administrations) are essential for high-stakes decisions. AI does not remove these requirements; it changes where risks originate and how they must be controlled. Increasingly, best practice recommendations emphasize human oversight, transparency, and continuous monitoring rather than one-time model certification.

Key Problems in AI-Based Exam Scoring.

Construct Underrepresentation and Construct-Irrelevant Variance. AI systems often model easily measurable proxies. In writing, length, vocabulary rarity, and syntactic complexity may inflate predicted scores even when task achievement is weak. In speaking, acoustic clarity may be conflated with proficiency; a candidate with strong language skills but a heavy accent or low-quality microphone may receive an unfairly low score.

Example (writing): An AI model might reward long essays with varied connectors ("moreover", "therefore") even if arguments are repetitive or off-topic.

Risk: The system measures stylistic signals rather than communicative effectiveness.

Example (speaking): A candidate's score drops due to ASR misrecognition of accented speech, which then propagates to content and fluency features.

Risk: Technical artifacts are converted into proficiency penalties.

Bias and Fairness Across Groups. Bias can enter through training data imbalance, label bias (human ratings reflecting stereotypes), or measurement bias (ASR poorer for certain accents). Fairness issues are particularly sensitive in multilingual populations because accent, socioeconomic factors (device quality), and access to preparation resources vary systematically.

Common fairness failure modes:

- **Accent-related score deflation** in speaking tasks due to ASR errors.
- **Topic and cultural loading**: prompts that favor certain backgrounds can create performance differences unrelated to language ability.
- **Differential feature validity**: the same observable feature (e.g., speaking rate) may correlate differently with proficiency across L1 groups.

Domain Shift and Generalization Failures. Models trained on one exam form, prompt type, or population may perform poorly on new prompts or different test-taker cohorts. Even minor changes—new topics, updated rubrics, different recording conditions—can shift the data distribution.

In high-stakes settings, such drift threatens score comparability across administrations, potentially undermining year-to-year equivalence.

Opacity, Explainability, and Contestability. Candidates, teachers, and policymakers often require understandable reasons for scores—especially for writing and speaking. Many AI models produce only a number and limited feedback. Without meaningful explanations, trust erodes, and legitimate errors are difficult to detect.

A key governance issue is **contestability**: Can a candidate appeal, and can the system provide evidence for or against the appeal?

Gaming and Adversarial Test-Taking. When scoring rules become predictable, candidates may optimize for the model rather than for genuine proficiency. In writing, this might include template memorization, keyword stuffing, or unnatural complexity. In speaking, candidates might manipulate pauses or pace. More recently, generative AI introduces additional threats: candidates may submit AI-generated essays or use real-time assistance. Even if proctoring exists, detection is imperfect, and policies differ by institution.

Privacy, Security, and Data Governance. Language exams collect sensitive data: voice recordings, writing samples, metadata, and sometimes identity documents. AI scoring increases data processing and retention, raising questions about:

- consent and purpose limitation,
- secure storage and access control,
- cross-border data transfer,
- reuse of responses for training.

Solutions: A Defensible Framework for AI-Assisted Scoring

Validity-First Assessment Design (Before Modeling). AI should be constrained by the construct, not vice versa.

Recommended controls:

- **Rubric operationalization:** define what the score means (e.g., task achievement, coherence, lexical control) and map each rubric dimension to observable evidence.
- **Task design for measurability without construct loss:** reduce prompt ambiguity, ensure speaking tasks elicit target functions (describing, arguing, negotiating) instead of only monologues.

- **Blueprinting and standard setting:** confirm score interpretations align with intended decisions (placement, certification).

Human-in-the-Loop Scoring and Moderation

Fully automated scoring is rarely defensible for high-stakes speaking/writing. A common solution is **hybrid scoring**:

- AI provides a preliminary score and flags uncertain cases.
- Human raters adjudicate: (a) borderline candidates, (b) low-confidence outputs, (c) fairness-sensitive groups, (d) appeals.

Operational mechanisms:

- **Double-scoring protocols** (AI + human; or two humans with AI as third reader).
- **Rater calibration and drift monitoring** to ensure human labels remain stable.
- **Discrepancy rules**: if AI-human difference exceeds a threshold, trigger review.

Psychometric Calibration and Ongoing Quality Monitoring

AI scoring must be embedded within a measurement system:

- **Calibration to a stable scale** (e.g., linking across forms).
- **Reliability evidence** (agreement statistics, generalizability).
- **Monitoring dashboards** for performance drift by prompt, site, device type, and demographic groups (where legally and ethically permissible).

Use routine "anchor responses" and periodically re-score samples to detect shifts.

Fairness Audits and Bias Mitigation

Fairness requires both measurement and action:

- **Disaggregated evaluation:** compare error rates and score distributions across accent/L1 groups, gender (if collected), regions, device classes, and socioeconomic proxies (carefully).
- **ASR bias mitigation:** accent-robust ASR, noise-robust preprocessing, and confidence-aware scoring that discounts unreliable transcripts.
- **Counterfactual checks:** ensure superficial features (essay length, microphone quality) do not dominate scoring beyond what the construct justifies.
- **Bias-aware training:** balanced sampling, reweighting, and evaluation against fairness metrics (e.g., differential prediction, equalized error where appropriate).

Explainability and Candidate-Facing Transparency

Provide explanations that are meaningful but do not enable easy gaming:

- **Rubric-aligned feedback** (e.g., coherence, support, vocabulary control) with examples from the response.
- **Confidence indicators**: when the system is uncertain, say so and route to human review.
- **Clear policy statements**: what data is used, how long it is stored, and how appeals work.

Security, Integrity, and AI Misuse Controls

To address AI-generated submissions and assistance:

- **Assessment redesign:** more in-person speaking, integrated tasks, and process-based writing (outline → draft → revision under supervision).

- **Secure browsers / proctoring** where appropriate, combined with privacy-respecting safeguards.
- **Forensics and anomaly detection:** detect suspicious patterns (e.g., sudden jumps, repeated templates), while minimizing false accusations.
- **Institutional policy:** define permissible vs. non-permissible AI use and align preparation practices with those rules.

Data Protection and Responsible Model Lifecycle

Adopt strong governance:

- data minimization, encryption, least-privilege access,
- vendor risk assessment (if third-party tools are used),
- model versioning and audit trails,
- documented incident response plans,
- explicit prohibition (or strict control) on reusing candidate data to train general-purpose models without consent.

Discussion

AI-assisted scoring is best understood as an engineering and governance problem inside a measurement framework. The central trade-off is between scalability and the risk of invalid or unfair score interpretations. The strongest argument for AI is operational: faster results, consistent baseline scoring, and resource allocation toward human review where it matters most. The strongest argument against uncritical automation is ethical and psychometric: language ability is socially embedded, and errors disproportionately harm candidates who already face structural disadvantages (accent stigma, limited access to technology, unequal preparation).

A realistic path forward is not "AI vs. humans" but "AI with accountable human oversight." Hybrid systems can improve consistency and throughput while preserving contestability. However, hybrid systems can also fail if governance is weak—e.g., if humans rubber-stamp AI outputs or if appeals are not meaningful. Therefore, exam providers should treat AI scoring as a continuously monitored system with periodic re-validation, not a one-time deployment.

Conclusion

AI-assisted scoring of foreign language proficiency exams offers clear benefits in scalability, speed, and operational consistency. Yet it also introduces substantial risks: construct underrepresentation, fairness issues across accents and groups, generalization failures, opacity, gaming incentives, and privacy concerns. The paper proposes a defensible solutions framework built on validity-first design, human-in-the-loop moderation, psychometric calibration, routine fairness audits, explainable reporting, integrity protections, and strong data governance. In high-stakes testing, the responsible position is not to maximize automation, but to maximize the defensibility of score interpretations and the fairness of decisions.

References:

1. AERA, APA, & NCME. (2014). Standards for Educational and Psychological Testing. American Educational Research Association.
2. Bachman, L. F., & Palmer, A. S. (1996). Language Testing in Practice. Oxford University Press.
3. Chapelle, C. A. (2001). Computer Applications in Second Language Acquisition. Cambridge University Press.
4. Kane, M. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

5. Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). Macmillan.
6. Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). Automated Scoring of Complex Tasks in Computer-Based Testing. Lawrence Erlbaum.
7. Xi, X. (2010). Automated scoring and feedback systems in language assessment: A review. Language Testing, 27(3), 291–311.
8. Zechner, K., & Evanini, K. (2020). Automated scoring of speaking tasks: Current trends and future directions. In The Routledge Handbook of Second Language Acquisition and Language Testing. Routledge.
9. OECD. (2021). AI in Education: Guidance for Policy Makers. OECD Publishing.
10. UNESCO. (2023). Guidance for Generative AI in Education and Research. UNESCO.