

IMPLEMENTING AUTHENTIC AI ASSESSMENT IN TESOL: CHALLENGES AND RESEARCH DIRECTIONS

Dilnoza Usmanova

Abstract

This article addresses the practical challenges of implementing authentic AI-enhanced language assessment in TESOL contexts. Drawing on a four-dimensional framework of authenticity, we identify key implementation barriers at technological, pedagogical, and institutional levels. We propose a research agenda to address these challenges and offer practical guidelines for TESOL practitioners navigating the integration of AI assessment tools. The article concludes with recommendations for interdisciplinary collaboration between language educators, AI developers, and assessment researchers.

Keywords: AI implementation, language assessment, TESOL, research agenda, educational technology

1. Introduction

Artificial intelligence technologies offer promising possibilities for language assessment, but their successful implementation requires addressing significant challenges related to authenticity. This article examines implementation challenges through the lens of a four-dimensional authenticity framework (contextual, interactional, consequential, and representational) and proposes research directions to address these challenges.

2. Current Implementation Challenges

2.1 Technological Challenges

Computational Resources: Truly authentic AI assessment may require substantial computing power not available in all educational contexts. Resource disparities may create inequitable access to high-quality assessment technologies.

Technical Integration: Implementing AI systems within existing educational technology infrastructure presents challenges. Many language programs use learning management systems with limited AI integration capabilities.

Data Requirements: High-quality AI assessment requires extensive training data. Smaller language programs may lack sufficient data for customization or validation.

Algorithm Transparency: The “black box” nature of some AI systems complicates validation against authenticity criteria. Educators may be unable to determine how assessment decisions are made.

2.2 Pedagogical Challenges

Assessment Literacy: Many language educators lack sufficient understanding of AI capabilities and limitations to implement these tools effectively.

Balancing Assessment Types: Determining appropriate roles for AI versus human assessment remains challenging, particularly for complex language skills.

Feedback Integration: Incorporating AI feedback into broader pedagogical approaches requires careful design to avoid overemphasis on machine-detectable features.

Learner Resistance: Some learners may resist AI assessment due to concerns about validity, fairness, or preference for human evaluation.

2.3 Institutional Challenges

Policy Development: Many institutions lack policies governing AI assessment use, raising questions about validity, accessibility, and academic integrity.

Staff Development: Professional development related to AI assessment implementation is often inadequate.

Cost-Benefit Analysis: Institutions struggle to evaluate return on investment for AI assessment technologies, particularly regarding authentic assessment outcomes.

Ethical Considerations: Privacy concerns, data ownership, and potential bias in AI systems raise significant ethical questions that institutions must address.

3. Integration Matrix: Current Status

The following matrix evaluates current implementation status across educational contexts:

Educational Context	Contextual Authenticity	Interactional Authenticity	Consequential Authenticity	Representational Authenticity
Higher Education	Moderate - some contextualized tasks	Low - limited dialogue capabilities	Variable - depends on implementation	Low - limited accommodation of diversity
Private Language Schools	Low - standardized assessments	Low - primarily one-way feedback	Variable - commercial pressures	Low - standard language focus
K-12 Settings	Low - often decontextualized	Low - limited interaction	Concerning - potential negative washback	Low - normative approaches

Self-Directed Learning	-	Moderate personalization	Low scripted interaction	-	Variable depends on learner attitudes	Low mainstream language models	-
------------------------	---	--------------------------	--------------------------	---	---------------------------------------	--------------------------------	---

This matrix highlights significant gaps in current implementation, particularly regarding interactional and representational authenticity.

4. Practical Implementation Strategies

4.1 Short-Term Strategies

Hybrid Assessment Approaches: Combine AI assessment with human evaluation, leveraging each for appropriate aspects of language performance.

Contextual Scaffolding: Provide rich contextual information around AI assessment tasks to enhance contextual authenticity.

Feedback Mediation: Train educators to help learners interpret and apply AI feedback within broader communicative contexts.

Transparency Practices: Clearly communicate to learners what AI can and cannot effectively evaluate, preventing misaligned expectations.

4.2 Medium-Term Strategies

Customized Implementation: Develop institution-specific frameworks for AI assessment integration based on learner needs and program goals.

Professional Development: Create comprehensive training programs addressing both technical and pedagogical aspects of AI assessment.

Assessment Ecosystems: Design complementary assessment approaches that collectively address all dimensions of authenticity.

Continuous Evaluation: Implement ongoing evaluation of AI assessment impact on teaching practices and learning outcomes.

5. Research Agenda

To address implementation challenges, we propose a research agenda organized around the four authenticity dimensions:

5.1 Contextual Authenticity Research

- Developing and validating context-rich assessment tasks compatible with AI evaluation
- Examining the relationship between contextual features and AI assessment accuracy
- Creating frameworks for adapting AI assessment to specific target language use domains
- Investigating multimodal integration in AI assessment

5.2 Interactional Authenticity Research

- Advancing dialogue-based assessment technologies that support authentic interaction

- Evaluating turn-taking and repair strategies in AI-human assessment interactions

- Developing metrics for evaluating interactional competence through AI
- Exploring the potential of LLMs for more contingent assessment interaction

5.3 Consequential Authenticity Research

- Studying washback effects of AI assessment on teaching and learning practices

- Investigating stakeholder perceptions and acceptance of AI assessment
- Examining transfer of learning between AI assessment contexts and real-world language use

- Developing approaches to enhance learner agency in AI assessment

5.4 Representational Authenticity Research

- Creating and validating AI systems that accommodate linguistic variation
- Developing assessment approaches for multilingual competence
- Investigating cultural bias in AI assessment and strategies for mitigation
- Expanding training data to represent diverse communication styles

5.5 Interdisciplinary Research Priorities

- Collaborative research involving TESOL practitioners, AI developers, and assessment specialists

- Mixed-methods approaches combining quantitative evaluation with qualitative insights

- Longitudinal studies tracking the impact of AI assessment implementation over time

- Action research by practitioners implementing AI assessment in diverse contexts

6. Case Study: Implementing Authentic AI Writing Assessment

To illustrate practical implementation, we present a case study of an English for Academic Purposes program implementing an AI writing assessment system:

Initial Challenges:

- System provided detailed feedback on grammar and vocabulary but limited feedback on rhetorical effectiveness

- Students focused primarily on sentence-level corrections rather than global improvements

- Faculty questioned alignment with program's genre-based writing approach

- System showed bias against non-standard expressions common in multilingual writing

Implementation Strategies:

- Created supplementary rubrics addressing rhetorical dimensions AI couldn't evaluate

- Developed faculty-led workshops helping students interpret AI feedback within genre expectations
- Implemented peer review focusing on content and organization to complement AI's linguistic focus
- Provided faculty training on guiding students to critically evaluate AI feedback

Outcomes:

- More balanced attention to both linguistic accuracy and rhetorical effectiveness
- Increased student agency in determining which AI suggestions to implement
- Development of metacognitive skills through critical engagement with AI feedback
- Improved faculty attitudes toward AI as a complementary rather than replacement tool

This case illustrates how thoughtful implementation addressing authenticity gaps can leverage AI benefits while mitigating limitations.

7. Conclusion

Implementing authentic AI assessment in TESOL contexts requires addressing significant technological, pedagogical, and institutional challenges. The proposed research agenda and implementation strategies provide a pathway toward more authentic AI assessment integration.

While current AI capabilities show varying degrees of alignment with authenticity dimensions, understanding these gaps enables more effective implementation. By approaching AI assessment as a complement to rather than replacement for human assessment, TESOL practitioners can leverage technological affordances while preserving the authenticity essential to communicative language teaching.

Future progress will require interdisciplinary collaboration between language educators, AI developers, and assessment researchers to create systems that better align with all dimensions of authentic assessment. This collaboration should be guided by clear pedagogical principles rather than technological possibilities alone.

References

1. Chapelle, C. A., & Sauro, S. (Eds.). (2022). *The handbook of technology and second language teaching and learning*. Wiley Blackwell.
2. Levis, J., & Suvorov, R. (2022). Automated assessment of second language pronunciation. In H. Mohebbi & C. Coombe (Eds.), *Research questions in language education and applied linguistics* (pp. 803-808). Springer.
3. Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256.

4. Ockey, G. J. (2021). An overview of AI and language assessment: Definitions, applications, and challenges. *Language Assessment Quarterly*, 18(2), 119-135.
5. Winke, P., & Isbell, D. R. (2022). The development, implementation, and ethical management of AI-based language assessments. *Language Assessment Quarterly*, 19(3), 231-240.

