Academic Journal of Education
ISSN: 3080-9940 (Print), ISSN: 3080-9959 (Online) | Vol. 3, No. 2, 2026

An Analysis of Paths for Integrating AI Assessment in English Teaching Scenarios

Zhiwei Qiu

Faculty of Social Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom

Abstract: This paper studies practical paths for applying artificial intelligence in assessment in English language teaching. AI-based assessment is defined here as the supervised application of automated scoring, speech or text analysis, adaptive task delivery, and generative feedback to support evidence collection and pedagogical decisions. AI will be used to build the assessment system; it will not function independently as an examiner. It is a staged path that begins with the construction of definitions and low-stakes formative uses, continues to teacher moderation and learner dialogue, and develops into institutionally regulated implementation with validation, privacy protection, fairness review, and appeal mechanisms. Examples of writing, speaking, reading, listening and classroom feedback are provided. The paper is theoretical and practical, and there are no original learners, datasets, experiments, scores or statistical results. Therefore, it is not expected that AI will improve accuracy, achievement, efficiency and fairness. At its core, the educational goals are to learn by example, to think independently, and to be accountable.

Keywords: AI-based Assessment; English Teaching; Formative Assessment; Language Testing; Human Oversight; Assessment Validity.

1. Introduction

Artificial intelligence is now often used to grade written answers, analyse speech, provide feedback and adjust tasks, and organise classroom data. Teachers regularly organise difficult performances under a time limit to attract the attention of English-teaching students. However, the speed of a system does not determine the educational value of its output. A numerical score, error label or fluent feedback message may be suitable for the purpose of evaluation and use of the result; otherwise, they are either irrelevant or misleading.

This paper refers to AI-based assessment as supervised application of automated scoring, natural-language or speech analysis, adaptive delivery, and generative feedback in the context of English assessment. No plagiarism detection, translation or unlimited chatbot use analysis is provided by these tools, and thus they have not been adopted in a particular evidence-based decision-making process. Integration is the organisation of AI functions in the curriculum, tasks, teachers' judgments, students' participation and institutional regulations, rather than adding them as a separate technological layer.

The discussion presents a path for formative classroom application, teacher-AI cooperation, multi-modal assessment, learner assessment literacy and institutional governance. It should also be feasible, reasonable, private, transparent and academically honest. Theories and Applications of the Paper. There are no original participants, school data, model outputs, test scores, experiments or statistical analyses. Therefore, the proposed pathways are design options and not verified achievements of AI in improving English proficiency, reducing teacher workload, or creating fairer assessment.

2. Assessments Before Automation

Any path for integration should start with the interpretation of what the evaluation aims to achieve. According to Messick, a general index of the extent to which interpretations and behaviours can be supported by scores is called validity [1]. Thus, a common mistake in technology projects can be avoided: choosing an available model first and then inventing an educational purpose. A school should specify the language ability, decision-makers, users, stakes and possible consequences of an AI function before using it.

The tasks for language assessment need to be related to people's language abilities and useful. Bachman and Palmer have organised the development of tests in accordance with the criteria of reliability, construct validity, authenticity, interactivity, impact and practicality [2]. AI scores can improve the regularity of operations, but if only easily observable external factors are rewarded, authenticity will be lost. At the same time, although there is a large amount of data for genuine-speaking tasks, it is not easy to process automatically due to problems such as accents, noise, interaction and lack of topic knowledge. Integration should consider all the attributes and not only maximise an easy-to-obtain index.

Assessment for learning offers a second support. Black and William believe that, to foster learning, one should provide feedback and organize various classroom activities [3]. Therefore, AI will only have a positive impact on the learner or teacher when it helps them make choices. The grammar label after the last grade may not be teaching material. Promptly have the learner compare the two sentence options, revise a paragraph, and explain the modifications; if this advice is correct and easy to understand, it will be more beneficial to their learning.

The two are production of information and the formation of a judgment. AI finds features, assigns a grade to them, learns patterns, and generates annotations. Teachers and students still need to determine whether the evidence meets the requirements of the desired ability, whether the environment has affected performance, and what response to take. Therefore, the system should be designed to support the process of obtaining evidence and not replace professional

judgment.

3. From Automatic Scoring to AI

Automatic evaluation in language teaching predates current generative AI. Research has explored the use of computer systems for automatic essay grading and raised problems such as the reliability, validity and application in teaching and psychometric analysis [4]. This history is relevant to address the present problems of the tool. Although the language model can generate more fluent feedback than the old scoring system, natural language does not guarantee that the judgment serves the purpose of construction or is backed by reasonable evidence.

A Good Evaluation System Should Consider Both the Scoring Model and the Overall Assessment System. Williamson, Xi and Breyer propose to evaluate automated scoring based on evidence of construct representation, association with human judgments, operational performance and consequences of use [5]. Agreement with a teacher in class is not the only kind. Teachers should also investigate whether people and machines focus on the same features, whether performance changes occur in different genres or for different groups of learners, and whether students adapt their writing according to the expected requirements of machines.

Professional Test Standards are also responsible for interpreting and applying the desired scores. The Standards for Educational and Psychological Testing cover validity, reliability and precision, fairness, accessibility, reporting and the new application of technology [6]. High-Stakes Decisions need more evidence than low-stakes behaviour. An experimental AI comment on a draft and an automated score for placement is not ethically the same. The latter needs written consent, guarantees, the supervision of qualified staff, and an effective appeal path for the results.

Research on technology and language assessment has advanced in both efficiency and innovation. Chapelle and Voss divide them into applications for improving existing processes and applications that create new evaluation scenarios [7]. Thus, planning is clear. Automatic transcription can simplify the review of oral portfolios and is mainly an efficiency function. Interactive simulations can motivate students to speak in new circumstances and are thus innovative. Each needs its own proof, and newness is not to be considered an effective way of communication.

4. Formative Classroom Path

A teacher-led formative cycle is a low-risk starting point. The teacher sets a specific learning goal, designs or chooses a task, determines what evidence is needed, and sets the role of AI. Identify potential cohesion breaks in the writing lesson by the system and raise questions about the audience or organisation. Do not give a final grade before using the scores have been verified. Learners are exposed to the suggestions, consider whether they are suitable, modify the suggestions, and discuss difficult problems with peers and teachers.

Thus, there will be no feedback loop. The first level is related to task completion and goals: whether the response can address the problem or fulfill the need of the user. The Second Layer is for Organisation and Discourse. Third, there are some language features to be mentioned. If the AI provides too many corrections at once, the learner will not be able to learn and may feel lost. Set priorities according to the goals of the lesson and save some problems for later teaching.

AI feedback should be considered a claim subject to review. A small window will display the related information, the reason for the recommendation, and one or more choices for the learner. Do not be too directive; instead, when dealing with problems of acceptability according to register or intended meaning, use a softer tone. Learners can keep a revision log of the original wording, AI suggestions, decisions and explanations. Log shows that conformity is the goal of language cognition and provides teachers with data on students' responses to correction.

Assessment of reading and listening can also be conducted with AI. The system can generate initial comprehension questions and suggest distractors, adjust the difficulty of practice, etc., but teachers need to verify their answerability, textual support, cultural offensiveness and language level. The generated questions may be too simple or contain content not in the original text. Therefore, they are task candidates and not pre-validated items. The teacher should save the original material and record significant modifications to the content of assessments.

5. Productive Skills Pathway

Writing assessment is a typical case because it is easy to submit and grade the text. A reasonable Design separates the characteristics that machines can measure from the general characteristics a course values. Spelling, sentence length, lexical repetition or organisational signals may be due to a specific individual, and they are not necessarily signs of poor writing ability. The purposes, audiences, contents, organisation and language control of the rubrics should be clear. AI outputs fill in the selected evidence fields, and then teachers evaluate the performance and attention to creativity, reasoning, source use, and contextual appropriateness.

Speech is also problematic. Automatic speech recognition and pronunciation analysis can support repeated practice, but the results may be affected by the quality of recording, speech speed, accent, code-switching, names, and background noise. A score outside the range of teacher observation will not be automatically modified. The ability to get along with other people in different circumstances is called interpersonal skills. A person's pronunciation index cannot be used to assess the quality of communication.

The Common European Framework of Reference for Languages (CEFR) has added descriptions of reception, production, interaction, mediation, online interaction and plurilingual or pluricultural competence to expand language education [8]. The descriptors can offer information for task design but are not algorithms themselves. If Artificial Intelligence will be used to conduct the assessment, then corresponding observations need to be made for each item. Teachers can learn from examples near the decision boundary, and if they have a written reason, will not follow the recommendation.

A Multimodal Portfolio is a practical alternative to single-automation. Collect drafts, recorded presentations, peer assessments, reflections and revisions of learning. AI can organise the evidence, transcribe speech and summarise repeated feedback, and students select artifacts and explain their development. Portfolio claims should be reasonable; a group may have shown a pattern and progress in a particular area but is not necessarily good in all other areas. Teachers can compare the interpretations of different classes and find inconsistencies in rubric application.

6. Learner and Teacher Agency

AI Integration Alters Assessment Literacy. Inform the learners about the data the system will and will not collect, how their data will be used, and who will access it. They need to know whether AI help is allowed in the task. Distinguish among formative help, disclosed cooperation and forbidden substitution in the rules. A general statement that all use of AI is cheating fails to consider legitimate tools, and unrestrained application may make it difficult to determine whose language ability a submitted product represents.

UNESCO's Beijing Consensus advocates for the enhancement of human capacities through AI in education and aims to empower teachers rather than substitute them; it also emphasizes inclusion and ethical, transparent, and auditable applications of educational data [9]. Therefore, a human-in-the-loop model is used for assessment. Set Goals, Collect Data, Spread Results and Handle Exceptions. Technology will reduce routine work; however, any expansion in responsibilities should be confirmed by inspection, as the results are not always consistent.

Combine language assessment and AI literacy in teacher training. The staff should learn about the model's errors, compare the interpretations of AI and humans, test prompts, discuss data protection, and practice explaining decisions to the learners. Do not have a training company conduct the training. A standalone example of a failure case can help teachers learn that the system's performance differs depending on the task, population, language, etc. Communities of practice can record recurring problems and revise local guides.

Learner agency also needs an appeal path. When AI gives a score or feedback that will affect a student's progress, the student should be able to request a manual review without penalty for disputing the system's assessment. Based on the original results, rubrics, machine-generated content and other circumstances, conduct an analysis. Institutions should state who has decision authority and how corrections are recorded. Otherwise, the nominally supervised system will be unable to be challenged in decision-making.

7. Institutional Governance Pathway

Classroom experimentation should not be expanded to institutional application without regulation. A school can start with an inventory of systems, functions, data flows, vendors, users and decision-makers. Every use case should have a responsible person and a reason for being. Risk is relatively high when the output determines the placement and promotion of students, when sensitive student data are processed, and when the system is not easy to audit. High-risk uses need to be independently verified or are not suitable for deployment.

UNESCO's Guidance for Policymakers points out that, when introducing AI in education, pilot tests, monitoring, evaluation, evidence collection, and fair and ethical planning should be conducted [10]. Therefore, the pilot should have pre-prepared questions instead of a general goal of innovation. Schools may be concerned about the inconsistency of the results in teachers' interpretation, harmful errors in feedback, different treatment of accents or language backgrounds, and whether learners understand the system. Pilot evidence will lead to a halt, revision or expansion.

The NIST AI Risk Management Framework is divided into the functions of govern, map, measure and manage for risk work, and attributes such as validity, reliability, transparency, explainability, privacy protection and fairness-aware bias management are identified [11]. Mapping in the context of English assessment and governance identifies learners, tasks, contexts and harms; measurement tests performance and limitations; management establishes controls, monitors, responds to incidents and determines retirement criteria. It is not a language-test validation certificate, and thus should not be used in place of assessment-specific evidence.

Generative AI is unstable, and different results may be obtained due to changes in the model. UNESCO's guidelines for generative AI advocate for human-centered, age-appropriate, ethical, safe, equitable and meaningful applications, as well as privacy protection and verification [12]. Institutions should save the model and interface version, prompt or template, date, etc., of the settings. Updates should trigger new checks and not be silent. Staff should not enter identifiable student work into public systems without the consent of the students, their parents or guardians, or other relevant authorities.

Procurement needs to specify the requirements for data ownership, deletion, security, access, model change notification, documentation, subgroup evaluation and exit plan. Claims of proprietary secrecy cannot be used to excuse the institution's decisions. If the vendor cannot provide sufficient proof of consequential use, the system will be limited to low-risk teacher support or excluded. Operational human supervision: Reviewers should have the time, ability, access to original performances and actual authority to alter the results.

8. Validation and Monitoring

First, the task results are directly linked to the AI's produced evidence, interpretations and decisions. The support materials can be experts' analyses of the task contents, comparisons with other students' self-assessments, consistency studies, analyses of other learner groups, and research on the effects. A Single Correlation or Accuracy is Not Sufficient. Evaluate the performance of all kinds of real-life genres and various proficiency levels and accents, devices, etc., that the system will be applied to.

Fairness review should consider all the data. An excellent general result may have issues in a small group. Institutions should use a reasonable way to conduct subgroup analysis that respects privacy and does not reduce identities to mere biological classifications. Qualitative analysis will also be carried out because teachers and students have been exposed to culturally insensitive feedback, mispronunciation of names, stigmatizing language, inaccessible interfaces, etc., which cannot be shown in statistics.

Continue to observe after deployment. The corresponding indicators are the proportion of teacher overrides, learner appeals, unresolved feedback errors, missing or corrupted recordings, system downtime, and post-update modifications. These are operating signals and not proof of accuracy. Reasons for an elevated proportion of overrides may include model decay, the introduction of a new task type, an ambiguous rubric or altered teacher demands. No cause has been found by investigation yet.

The Report has errors and omissions. Learners should know if a teacher has given any feedback, recommendations or confirmations of the results. Score reports should not have false positives due to short-term performance sampling. Distinguish between classroom feedback and certified

proficiency claims at the institution. If the evidence is lacking, one should not speak confidently about compensation; instead, limit the use of the system, gather more evidence, or suspend the system altogether.

9. Summary

Integration of AI-based assessment in English teaching should be regarded as assessment redesign, not software installation. A reasonable road begins with the creation of language and an intended choice; next is low-stakes formative support to foster teacher-and-learner assessment literacy, and only after verification and governance will it be widely spread. AI can help collect evidence, recognise patterns, draft tasks, transcribe, etc., in the areas of writing, speaking, reading, listening and portfolios. It is not to be mistaken for the test of language ability.

Norms have various functions in practice. Machines can process recurring features and organise information; teachers interpret performances in the curriculum and society; learners learn from feedback and keep a record of questioning decisions; institutions manage privacy, fairness, procurement, monitoring and appeals. Human supervision is only effective if the supervisor has knowledge, time, evidence and authority.

This paper has put forward some ideas but has not conducted an experiment. There is no original dataset, participants, model evaluation, scores or measured results. Therefore, it is not claimed that AI will necessarily enhance accuracy, efficiency, achievement, motivation and fairness. The end needs some specific experimental data. Until such evidence is available, the English program should be of a small scale, reversible in nature, limitations should be disclosed, and it should be avoided if the educational objectives or risks cannot be addressed adequately.

References

[1] Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). Macmillan.

[2] Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice: Designing and developing useful language tests. Oxford University Press.

[3] Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74. https://doi.org/10.1080/0969595980050102.

[4] Shermis, M. D., & Burstein, J. (Eds.). (2003). Automated essay scoring: A cross-disciplinary perspective. Lawrence Erlbaum Associates.

[5] 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. https://doi.org/10.1111/j.1745-3992.2011.00223.x.

[6] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

[7] Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. Language Learning & Technology, 20(2), 116–128.

[8] Council of Europe. (2020). Common European framework of reference for languages: Learning, teaching, assessment—Companion volume. Council of Europe Publishing.

[9] UNESCO. (2019). Beijing consensus on artificial intelligence and education. UNESCO.

[10] Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). AI and education: Guidance for policy-makers. UNESCO. https://doi.org/10.54675/PCSP7350.

[11] Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1.

[12] Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO.