

- bilingual reading modes: Human-AI robot interaction and paper books. *Frontiers in Psychology*, 14. <https://doi.org/10.3389/fpsyg.2023.1200675>
- Fu, Y., & Hiniker, A. (2025). Supporting students' reading and cognition with AI. *arXiv*. <https://doi.org/10.48550/arXiv.2504.13900>
- Hockly, N. (2023). Artificial intelligence in English language teaching: The good, the bad and the ugly. *RELC Journal*, 54(2), 445–451. <https://doi.org/10.1177/00336882231168504>
- Kessler, G., & Plakans, L. (2008). Does teachers' confidence with CALL equal innovative and integrated use? *Computer Assisted Language Learning*, 21(3), 269–282. <https://doi.org/10.1080/09588220802090303>
- Liu, Z. (2005). Reading behavior in the digital environment. *Journal of Documentation*, 61(6), 700–712. <https://doi.org/10.1108/00220410510632040>
- Moorhouse, B.L., & Wong, K.M. (2025). *Generative artificial intelligence and language teaching*. Cambridge University Press. <https://doi.org/10.1017/9781009618823>
- Rukiati, E., Wicaksono, J. A., Taufan, G. T., & Suharsono, D. D. (2023). AI on learning English: Application, benefit, and threat. *Journal of Language, Communication, and Tourism*, 1(2), 32–40. <https://doi.org/10.25047/jlct.v1i2.3967>
- Shafiee Rad, H. (2025). Reinforcing L2 reading comprehension through artificial intelligence intervention: Refining engagement to foster self-regulated learning. *Smart Learning Environments*, 12, Article 23. <https://doi.org/10.1186/s40561-025-00377-2>
- Wolf, M. (2018). *Reader, come home: The reading brain in a digital world*. Harper.
- Woo, J. H., & Choi, H. (2021). Systematic review for AI-based language learning tools. *Journal of Digital Contents Society*, 22(11), 1783–1792. <https://doi.org/10.9728/dcs.2021.22.11.1783>
- Zafar, N., Saira, & Afzal, S. (2025). AI-powered reading support for multilingual learners in higher education: A critical review. *Journal for Social Science Archives*, 3(1), 776–786. <https://doi.org/10.59075/jssa.v3i1.158>

Appendix A: Sample Data Collection Instruments and Raw Data Excerpts

1. Semi-Structured Interview Guide (Sample Questions)

How frequently did you use AI tools during your out-of-class reading assignments this semester?

Which specific AI features (e.g., summarisation, vocabulary explanation, paraphrasing) did you find most helpful for understanding academic texts?

Can you describe a time when you disagreed with or questioned an output provided by AI?

In what ways, if any, do you feel that using AI tools changed the way you approach a difficult reading passage?

Did you ever feel that relying on AI made you less likely to read the text closely? Please explain.

2. Sample AI Interaction Log (Participant: Lale)

User Prompt: I am reading a text about climate change. Can you explain the following paragraph like I am a high school student? [Text of paragraph pasted]

AI Output: Sure! Basically, this paragraph is saying that while the earth has always had temperature changes, the current warming is happening much faster because of human activity like burning coal and oil...

3. Sample Learner Journal Entry (Participant: Mehmet, Week 5)

Today I felt slightly tired and the text about Economics was very hard. I used ChatGPT to summarise the whole thing before I even tried to read it myself. Later, I realised I didn't actually learn any of the new words from the text. I felt lazy, so next time I will try to read first and use AI only for the parts I don't understand.

Ferit Kilickaya is a Professor of English language education specialising in technology-enhanced learning, assessment, and TESOL research at Burdur Mehmet Akif University in Türkiye.

ferit.kilickaya@gmail.com

English Language Assessment and Feedback Using Tech-driven and Generative AI Tools: New Directions for Teachers

Lina Mukhopadhyay

Abstract

This paper argues for building teacher agency in the use of tech-driven and generative AI tools for classroom-based language assessments. To attend to this aim, the facilitative role of AI tools to design language assessments and provide feedback are presented through a critical review. Current research directions of the impact of using such tools in English classroom assessments are presented for teachers to understand how to employ the tools for designing and administering language assessments, creating estimates of learning to provide feedback and helping learners easily identify recurring language learning difficulties, all the while focusing on creative and communicative use of English as a Lingua Franca. The paper concludes by drawing implications and challenges of using tech-driven and AI tools to complement teacher-led assessment practices and make language education inclusive and democratic.

Keywords: English language assessments, generative AI tools, tech-driven tools, CEFR-aligned rubrics, teacher agency, English as a lingua franca

Introduction

Language assessment for low-stakes teacher use in the English language classroom context requires to be localised to learner proficiency needs. The purpose of such assessments is to prepare learners perform well in a variety of communicative contexts ranging from academic to social

and professional. Various challenges abound teachers as they have to routinely deal with large cohorts of learners in Indian classrooms who have heterogeneous levels of general English proficiency and comprehension skills. Furthermore, they are often riddled with a bias of considering only standard English and not English as a Lingua Franca (ELF) (Jenkins, 2012; Jenkins et al., 2018).

Designing valid assessments and providing feedback often becomes daunting for Indian teachers as they hardly ever receive any professional training on how to deal with these parts systematically. Additionally, catering to large learner groups in classes coupled with heavy workload are not always conducive to design classroom assessments and use them reliably. In such a scenario, technology-driven language analysis tools and generative Artificial Intelligence (AI) tools can prove to be potentially powerful to help teachers design and use classroom assessments in valid and reliable manner. These tools can be leveraged to build teacher agency in the model of language assessment *for learning* (Earl, 2013). This would involve increasing teacher knowledge and skills to generate language items using AI tools (Shin & Lee, 2024), understand the ethical use of AI generated materials (Cong-Lem et al., 2024), and prepare personalised CEFR-aligned (Council of Europe, 2001, 2020) feedback on learner performance. To give an example, at lower levels, the tools can be used to identify short texts at A1, A1+ and A2 levels with comprehension questions to identify factual details and present short oral or written discourse to express everyday events, while at higher levels of B1 to B2 levels and above, it can create texts with higher complexity and add questions on inference generation and critical thinking, and attend to academic style, coherence, and fluency in production (Barrot, 2023).

Teachers need training in valid and ethical use of these tools (da Veiga, 2025) and leveraging the potential use of the tools in giving consistent and continuous feedback to make learners become more self-regulated (Shin, 2023). The aim of this article is to present contemporary research directions from English classrooms globally (Chuang & Yan, 2025) and provide strategies to help Indian teachers understand the complementary role of technology-driven language software and generative AI tools in assisting them to design and use language assessments (further reviewed in Shin et al., 2025). The complementarity of the tools would be presented through relevant and current research directions. Furthermore, suggested steps to design and administer language assessments, to create

estimates of learning to provide feedback and to help learners easily identify recurring language learning difficulties along with focusing on the creative and communicative use of ELF (Jenkins et al., 2018) across different global contexts will also be considered.

The article is organised in the following manner: it begins with a discussion of the construct of teacher agency in using tech-driven and generative AI tools for classroom-based language assessments followed by the different aspects of use of the tools for various parts of language assessment design. It concludes by drawing implications and challenges of using the tools to leverage agency in teacher-led assessment practices by democratising English use in classroom context (Tafazoli, 2024) as well as its use as a lingua franca (Jenkins, 2012).

Teacher Agency in Design and Use of Language Assessments

Language teaching and assessment methods have evolved overtime by being impacted from theorisations and empirical findings from classroom-based and experimental research in applied linguistics and pedagogy. Research on teacher use of generative AI tools and their learning affordances is still at a nascent stage. The gains along with the dilemmas and lack ethical guidance in valid and reliable use of the tools to support language learning in class is just developing (Barnes & Tour, 2025). Likewise there is scant research to show the uptake of teacher use of such tools to design language assessments and draw valid conclusions about learner progress in Global South contexts like in India. However, teacher agency as a construct in the success of ESL/EFL learning in the instructional context is undeniable and has empirically proven to be a sound practice globally (Priestly et al., 2015; Cong-Lem, 2021). Thus, at a time when tech-driven and generative AI tools are available freely and are open access, there is an urgent need to raise teacher awareness and build teacher agency in using these tools realisably and ethically. Research to examine teacher knowledge and classroom practices in using AI tools creatively to assess learners and scaffold learning is also required. Hence, in this paper I argue for building teacher agency by presenting strategies through which they can leverage tech-driven and generative AI tools to design and use assessments for formative purposes and thereby create opportunities for ‘assessment *for* learning’ (Earl, 2013) and provide ‘transformative feedback’ (Popham, 2011) to scaffold language learning.

Using Tech-driven and Gen AI Tools for Language Assessment

The use of tech-driven and generative AI tools in ESL/EFL assessment can be used to democratise English proficiency assessment from one standardised form of academic use of the language to assessing English as ELF (Cogo et al., 2021) with its varied forms and registers to support localised language assessments in different global regions where English is only expected to be used for academic purposes. Some of the widely recognised AI-powered chatbots are OpenAI's ChatGPT (GPT-4.5) (Zhu et al., 2023), Microsoft's Copilot, and Google's Gemini (v2.5), that can give virtual assistance in step-by-step design of language assessments, rubrics and personalised feedback plan. But the quality of all of these components would be based on well-crafted personalized teacher prompts for each specific aspect of an assessment tool. In what follows is a critical review of research, strategies along with challenges for teachers to employ tech-driven and generative AI tools for designing culturally sensitive ESL/EFL assessment.

Text Selection and Generation

Differentiated content selection and text simplification are useful strategies in offering individualised support in language learning. The use of differentiated texts in design of language assessments can be central to selecting texts and design reading and other integrated skills assessments. Teacher ability in selection and use of differentiated texts to match learner levels of proficiency and needs has been reported to be useful and reflective (Liu et al., 2024).

A tool for checking the level of texts aligned to the CEFR (Council of Europe, 2001, 2020) levels is 'readability checker' by cathoven 'language hub' meant for teachers (<https://www.cathoven.com/readability-checker/>). This is available for free and can be used to assess the level of a text by assessing vocabulary coverage, grammar and discourse level difficulty of the text. As a second step, another function from the same tool can be used called 'level adaptor', which can adjust the texts to learners' required level of vocabulary and sentence structure without altering the key ideas of the texts (<https://www.cathoven.com/level-adaptor/>). Furthermore, teachers can also use ChatGPT or Copilot to effectively assess and reduce (or increase) the linguistic complexity of text materials while retaining the original information (Alva-Manchego et al., 2020) and/or generate novel texts in keeping with the levels of

CEFR (Uchida & Negishi, 2025). The language of AI generated texts would have to be modified to make it authentic and culturally relevant.

Vocabulary and Grammar Inputs

As assessment of vocabulary and grammar are integral to English as a second/foreign language (ESL/EFL) assessments particularly at Beginners' and Intermediate levels, corresponding to A1, A2 upto B2 levels of proficiency of CEFR (2001 & 2020), teachers can use Grammarly, a free AI writing assistant to help learners get individualised feedback on their language production. However, as research shows, teachers would need to intervene by observing the changes in writing (or speech) in the sub-areas of grammar where the learners would have divergent interlanguage output in their previous attempts (Han & Li, 2024). They would need to raise learner awareness that grammar correction is not for isolated items per se but to improve overall discourse coherence as well as attend to the clarity of expression. Tools like ChatGPT and Gemini can also help learners gradually increase structural complexity by showing how to use subordination to express links between ideas cohesively.

A useful tool to analyse vocabulary quality on the measures of 'lexical breadth' (number of new words used in a text) and 'lexical depth' (the use of same root for variety of grammatical functions and/or meaning thorough inflectional and/or derivational morphology) is lexical tutor. It is a web-based data driven vocabulary analysis tool available for free of cost. In it, there is a function *vocabulary profile* (<https://www.lextutor.ca/vp/>) using which any text can be analysed for the measure of 'lexical quality' (breadth and depth). The profiler categorises vocabulary into British National Corpus (BNC) frequency starting from K1 level (or first 1000 frequently occurring words in British English) up to higher levels of K20 and above. For instance, usually any text for primary grade Indian learners (till about grade five) should have 90 per cent of the words in the K1 to K2 level for the texts to be adequately comprehended (by attaining 60 per cent to 70 per cent success in comprehension) (Laufer, 2020). Using this tool, teachers can analyse the lexical quality of texts to be used in primary grade reading assessments as well as the production knowledge of their learners. In all, using the tools of grammar and vocabulary, the teachers can also plan to provide level appropriate individualised feedback to the learners (Barnes & Tour, 2025).

Designing Authentic Items/Tasks and Rubrics

In assessing language skills and elements (grammar and vocabulary) formatively, teachers need to design ESL/EFL items and tasks that are authentic or close to real-life use of English for communication in the local context of use along with expression of academic knowledge and discourse using the standard English register. However, unlike summative assessments, teachers need not be bound by one strict formal register of English but can consider ELF versions (Jenkins, 2012; Jenkins et al., 2018) to accommodate non-standard features of grammar, vocabulary and discourse as used locally in Global South contexts like Indian classrooms. Teachers could be encouraged to use these instances as a benchmark of successful performance rather than compare learner performance against standard (and often commercialised) forms of the target language assessment (Harding & McNamara, 2018).

To design authentic ESL/EFL assessments that can encourage the usage of ELF, teachers could select language from Indian newspapers and magazines and use ChatGPT, Copilot and Gemini to rephrase or summarise texts. They can also use the tools to design items and tasks that could range from fixed-responses to free-responses. However, they would need to be critical of the quality of items designed by the AI tools. They would need to modify and revise the items to address ambiguities in ideas and adjust the level of language complexity, making the items reliable for classroom use. Although ChatGPT, Copilot and other tools like [revisely.com](https://www.revisely.com) or [mentimeter.com](https://www.mentimeter.com) can be used to generate language quizzes and other tasks, the validity and reliability of those have to be verified by the teacher's sense of judgment. Without teacher modifications, the tools can run the risk of being faulty and therefore produce performance results that may not be correct or reliable and give rise to negative washback.

Research on teachers using ChatGPT-generated language items have garnered attention in the recent times. Shin and Lee (2024) note that the rapid adoption of tool generated items by human item writers is an optimistic trend in classroom language assessment; but Shin et al. (2025) also caution that there is an urgent need of empirical evidence of validating AI generated items. Once the items are generated, they advise that teachers would necessarily have to check the item quality for their thematic flow, alignment to the content, and areas covered for assessment (Shin & Lee, 2024).

Furthermore, AI driven tools like ChatGPT, Gemini and Copilot can be used to design rubrics for language assessment that are aligned to specified CEFR level descriptors (2001, 2020). A recent review of research on ChatGPT to create language rubrics and assess writing using such rubrics by Uchida and Negishi (2025) shows that the tools have prior knowledge of CEFR descriptors. This knowledge they argue can be leveraged, such that teachers can potentially use the AI tools to function as reliable automated writing evaluators complemented by their moderation to assess written (or oral) performance.

However, a note of caution, for designing rubrics, teacher prompts have to be aligned to the task at hand. Furthermore, after rubric generation teachers would have to closely revise and modify these rubrics to localise them according to the perceived levels of learners use of ESL/EFL for academic activities and social communication and incorporate features of ELF. Thus, teacher knowledge and creativity would be central to using the tools to design items, tasks, and rubrics of ESL/EFL assessments to suit local contextual needs (Uchida & Negishi, 2025).

Summarising Results from Language Assessments

Language assessments once administered are either scored with binary options for fixed-response such as multiple choice question (MCQ) items or evaluated on rubric-based letter grades (or equivalent scores) for free-response or open-ended tasks. The scores and grades or both as performance indicators are available for multiple purposes. The primary function of course would be to inform the learners about their performance and convey through feedback the quality and process of language learning. Additionally, numerical information available through scores can also be summarised to get group estimates and make comparisons, between and within groups, longitudinally for an academic year. Subsequently teachers can summarise group performance and communicate it to the relevant stakeholders as the learners, parents and administrators to make the assessment process transparent and reliable. Teachers can also gain insights into the quality of assessment tool design and learner uptake, based on which they can make plans of adjusting pedagogy and assessment practices.

To obtain reliable summary of group performance, teachers can employ an AI tool like ChatGPT to get output in the form of descriptive statistics such as central tendency (mean, median and mode), dispersion (range,

standard deviation), percentage scores, normality of group performance, standard error and histogram. The formula for each of these functions to be used in Excel sheet can be generated from ChatGPT. Alternatively, Copilot can generate statistical output and diagrams of group performance. For this, the first step would be to provide the details of the group such as number of participants and their scores in each test or assessment and total score in test/assessment. Once the statistical output is generated in a tabular format it would be available for downloading as a csv file while the diagrams (as box plot, whiskers plot, line graphs and charts) can be downloaded as images. These can be incorporated as findings to be reported as directions of learner performance that a teacher might like to prepare for administrative or research or both purposes.

However, there is a point of ethical concern. Although AI tools like ChatGPT and CoPilot are very efficient in producing statistical results and might seem to decrease teacher workload to provide efficient summary estimates of learner performance very fast, teachers need to have awareness on the ethical use of data analysis (Schwartz, 2025) for language assessment purposes. For instance, ChatGPT or Copilot can wrongly prompt the teacher if she wants to generate simulated results. Teachers should be cautioned and made aware that simulated results are only an example to understand how to interpret trends and should never replace real data sets of learner performance. This would be unethical and have negative implications both for classroom use as well as for research. Hence, the tools can be efficiently used to statistically summarise performance with the awareness that simulated results are unethical to use.

Providing Personalised Feedback

Evaluation of learner performance on language assessments used for formative purposes would not be complete without providing feedback that could potentially transform the language learning process (Popham, 2011). Feedback when meaningful, constructive and aligned to the target language needs of the learners would make them more engaged to the process of learning. The use of tech-driven and generative AI tools can not only help teachers draw estimates of performance but also help them in preparing feedback that is constructive to scaffold learning (Shin & Lee, 2024; Barrot, 2023). Furthermore, learners can be trained to use the tools

to self-assess their language use ability and get self-feedback, thereby making them more agentic. It would create opportunity for assessment *for*, and *as* learning, instantiating self-regulated learning (Earl, 2013).

A strategy to generate feedback on writing performance could be one where the teacher can collect pieces of learner compositions and use the rubric, previously generated by the AI tools and teacher modified (Barrot, 2023), to identify the recurrent themes in learner texts. Thereafter, she could interpret the quality of writing vis-à-vis the themes and give content feedback to the learners, either in small groups or individually. This could be an instance of ChatGPT-supported teacher feedback. Han and Li (2024) have demonstrated that such feedback has proven to be beneficial to improve rhetoric of writing, attend to a diverse range of error categories, and make learners more engaged to their writing corrections, leading to long-term improvement of writing quality.

The potential of ChatGPT as a language learning tool can also be leveraged through self-assessment where teachers can ask intermediate level and above learners to critically document their process of composing and revising their drafts as pre-, while- and post- writing with the help of the tool and how they benefitted from the feedback they received from the tool as reviewed by Barrot (2023). It has the possibility to become an adaptive L2 writing assistant tool that is readily accessible and has an easy conversational and non-intrusive method of providing feedback. However, as an advisory, teachers need to raise learner awareness of ethical practices of using AI tools to ensure the integrity of writing ability as developmental and not a product (van Dis et al., 2023). Furthermore, they need to alert learners that the tools can also produce inaccurate results for content development as well as have rigidity of templates and language use limiting their use for assessment and providing feedback on learner texts. A limitation identified by Barrot (2023) is that the tools may generate the same set of rubrics for texts that are thematically very different but analyse them with similar surface level structural similarity of ideas.

Classroom-based Research on Usefulness of Integrating Gen AI in Language Assessments

The affordances provided by the use of tech-driven and generative AI tools in language assessments for classroom purposes can be verified and validated by teachers if they engage in small-scale classroom-based

research and share their experiences and findings with other colleagues. One of the areas which would be interesting to undertake is teacher-led classroom research to examine the reliability of AI constructed English language assessment items (for a full review refer to Shin & Lee, 2024), which are based on a statistical deep language modelling approach (Radford et al. 2019). They can also check the validity of the 'proficiency construct' in terms of the (a) language abilities, (b) context of use and (c) level(s) considered to generate the items by preparing suitable prompts. Post item creation, the AI tool can be prompted to share the details of item construct. At this stage a teacher can compare her estimates of the items based on her prior knowledge of the language level(s) suitable for her learners and modify the items. For instance in a study by Shin (2023) the efficacy of automated reading item generation by ChatGPT was reported where teachers found the items to be efficient and convenient but with a caveat that the prompt for each question type (e.g., vocabulary in context, character's intent/mood) needs to be carefully crafted. Hence, this kind of an in-depth reflective approach to understand the validity of constructing language items/tasks would help teachers understand the usefulness alongside the limitations of the tool(s). It would help teachers learn when to intervene and the need to review and revise items (Shin & Lee, 2024; Shin et al., 2025).

A second area of investigation would be teacher experience with planning individualised feedback using the AI tool(s). Reflective understanding of how AI tools identify themes in content development and summarise trends for each learner would add to teacher professional development. When teachers would be able to notice which aspects of feedback—content, language, or style of writing—the tools are able to provide, they would be better prepared to deliver feedback and feel confident in doing so. To complement AI-generated points, the teacher can plan which aspects she needs to add further inputs on and/or explain to the learners how to incorporate the feedback to improve without becoming heavily reliant on the tool generated language or content. Another area of reflection could be to notice and document how the learners have worked with the feedback and how they have shown growth in writing proficiency over time. Such observations can also be shared with the learners to further build their agency in incorporating feedback and develop proficiency in a self-regulated manner.

Conclusion

Currently there is widespread use of technology driven and generative AI tool-based strategies for assessing ESL/EFL learners globally. This gained impetus in the wake of the pandemic years, and have provided considerable affordances. Generative AI tools have been found to have considerable transformative potential in providing localised and individualised professional development opportunities for teachers and is democratising English education by tailoring learning experiences of teachers with diverse cultural backgrounds and learning needs (Tafazoli, 2024). But it has also brought challenging issues to the fore such as the validity, reliability and authenticity of teacher-led localised English language assessments tools and feedback providing methods (Lobalsamo et al., 2024; Chuang & Yan, 2025). Hence, to use the tools for language assessment for classroom purposes, teachers need to be agentive and responsible in evaluating the efficacy and usefulness of the tools and their impact on classroom use. Teachers can raise their awareness of this growing field of AI-driven language assessments by reading related research as presented in this paper as well as experiment with the tools to understand their scope of use. Alongside this, teacher agency can be built by training them to follow the ethical use of the tools to preserve academic integrity of language assessment and scaffold learning, as this is a growing concern in academia and research.

References

- Alva-Manchego, F., Scarton, C., & Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1), 135–187. https://doi.org/10.1162/coli_a_00370
- Barnes, M., & Tour, E. (2025): Teachers' use of generative AI: A 'dirty little secret'? *Language and Education*, DOI: 10.1080/09500782.2025.2485935
- Barrot, J.S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, 100745. <https://doi.org/10.1016/j.asw.2023.100745>
- Cathoven AI. (2025). *Cathoven* [AI Text Analyser and Generator]. www.cathoven.com
- Chuang, P., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. *System*, 134. <https://doi.org/10.1016/j.system.2025.103846>
- Cobb, T. (n.d.). *Lexical tutor*. <https://www.lextutor.ca/>
- Cogo, A., Fang, F., Kordia, S., Sifakis, N., & Siqueira, S. (2021). Developing