
TEACHERS' FORUM

Principles and Practices for Integrating genAI in Formative Language Assessment

LARRY GRULLON-POLANCO

University of South Florida

E-mail: larrygrpolanco@gmail.com

JIHYE KIM

University of South Florida

E-mail: jihyekim1@usf.edu

Recent developments in generative artificial intelligence (genAI) have ushered in a shift in the educational paradigm. Despite the potential pitfalls in genAI, research in this field has also indicated that these technologies possess the capability to improve teachers' workflows and have broader impacts on educational accessibility and effectiveness (e.g., Sullivan et al., 2023). Nevertheless, a significant gap persists in practical, in-class applications and clear, principled guidelines for integrating these technologies into everyday teaching. This article seeks to provide actionable strategies and prompts to ease teachers' workloads and improve teaching by demonstrating how ChatGPT can generate formative language assessment items. In this investigation, we utilized three of TESOL International's principles for exemplary teaching to develop formative language assessment items. By aligning genAI technology with established pedagogical strategies, this research highlights the role of genAI as a supplementary tool rather than a substitute for human educators.

INTRODUCTION

Advancements in generative artificial intelligence (genAI) are reshaping education. These tools utilize powerful underlying models, such as Large Language Models (LLMs), which learn patterns from vast training datasets to generate content ranging from text and images to music and video. They have the potential to offer teachers and learners worldwide a powerful tool to improve workflow efficiency, streamline lesson planning, and enhance instructional practices (Berg & Plessis, 2023). Furthermore, these technologies are expected to transform how learning is facilitated and accessed (Zawacki-Richter et al., 2019).

Given genAI's rapid development, researchers in educational technology and language education have conveyed concerns and optimism regarding the implications of their integration in educational settings (e.g., Dillenbourg, 2016; Giannakos et al., 2024; Thorp, 2023). While several inquiries remain unanswered (e.g., so-called "hallucinations," namely, inconsistent outputs) (e.g., Kooli et al., 2023), scholars have highlighted the importance of the ethical use of genAI (e.g., Giannakos et al., 2024) and the role of teachers (e.g., Dillenbourg, 2016) to leverage the benefits of these technologies effectively. Although valid concerns regarding genAI potentially replacing teachers remain an issue that educational institutions need to thoughtfully address, scholarly discourse is increasingly shifting to examine how genAI

can enhance teaching and how human and genAI strengths can complement each other (Carter & Nielsen, 2017; Hrastinski et al., 2019; Minnillo et al., 2024).

A growing number of studies in education are exploring how teacher-AI collaboration might improve the overall learning experiences of students (e.g., Holstein & Alevan, 2022; Ji et al., 2023; Qin et al., 2020). At the same time, many educators are already stretched thin. They face increased stress, anxiety, and burnout due to the significant portion of their time consumed by non-teaching tasks (Higton et al., 2017; Agyapong et al., 2022). In this regard, using genAI as teaching assistants can lighten the workload by assisting in the preliminary stages of lesson planning or materials development.

While previous studies have provided valuable insights into the collaboration between teachers and genAI in education, less attention has been paid to the specifics—the what, when, where, and how—of working with genAI to directly support teachers in their pedagogical tasks. The effective use of genAI in language assessment depends on its integration with thoughtful pedagogical planning. Additionally, while most teachers agree that genAI will have a significant effect on language education, many of them also express a lack of competence in using it (Moorhouse & Kohnke, 2024). Educators play a vital role in guiding these technologies, evaluating their outputs, and ensuring alignment with learning objectives.

Thus, this paper describes the incorporation of genAI into teaching and learning practices aligned with TESOL International's 6 Principles (Hellman et al., 2019). While genAI encompasses models that produce varied content, our focus is on LLMs, which are trained on extensive linguistic data to generate human-like text and perform language-related tasks (e.g., responding to inquiries, offering examples, etc.) (Kansneci et al., 2023). LLMs have attracted considerable attention in the fields of education and applied linguistics, as they have shown potential to serve as learning tools or resources for students (e.g., Du & Hashimoto, 2025; Wang & Reynolds, 2024). Specifically, we utilize ChatGPT, developed by OpenAI, as it is one of the most widely recognized LLMs; numerous studies have already explored the perceptions and acceptance of ChatGPT (e.g., ElSary, 2023; Habibi et al., 2023; Yilmaz et al., 2023). By choosing a tool with which teachers may have some familiarity, we lower potential barriers to engagement and ensure our findings are more accessible and readily applicable. To do so, this paper demonstrates how to generate formative language assessment items using principled prompts, making the case for how genAI can effectively serve as a supplementary tool rather than as a substitute for teacher expertise.

In the section below, we situate our work in the context of existing scholarship on AI-assisted language learning. We then provide an overview of theoretical insights provided through TESOL International's 6 Principles (Hellman et al., 2019) and recommendations regarding language assessment by Hughes and Hughes (2020), before providing a specific prompt template teachers can employ in their teaching practice. We conclude with a brief demonstration of implementing principled genAI language assessment in an example case.

AI-ASSISTED LANGUAGE LEARNING

AI-assisted Language Learning (AILL), as an emerging trend, has focused on integrating various AI technologies into language learning and teaching (e.g., An et al., 2023; Chun, 2022; Du & Hashimoto, 2025; Huang et al., 2022; Minnillo et al., 2024; Wang & Reynolds, 2024; Weng & Chiu, 2023). A substantial body of research represents diverse applications of AI technologies in language learning and teaching, ranging from automated language assessment tools to intelligent tutoring systems. This section provides a selective review of recent studies

focused solely on LLMs, such as ChatGPT, as they bear on language learning and teaching. Previous studies have examined various aspects of LLMs, including their affordances (e.g., Ahn, 2023; Baskara, 2023; Peres et al., 2023; Kohnke et al., 2023), their role in materials development (e.g., Shin & Lee, 2023; Shin, 2023), and students' perceptions and attitudes toward these technologies (e.g., Liu & Ma, 2024).

Research on applying LLMs in language learning and teaching has demonstrated considerable potential in promoting successful language learning outcomes. For instance, Carlson et al. (2023) conducted a study investigating the utilization of OpenAI's GPT-4 for providing feedback on English learners' writing tasks in English for Academic Purposes (EAP) classes. Research has also begun to address the need to incorporate genAI literacy for students within similar EAP contexts (e.g., Ngo & Hastie, 2025). The authors argue that when compared to GPT-3, GPT-4 demonstrates enhanced reliability and an increased capacity to interpret more nuanced instructions due to its larger model size and the extensive training dataset from various internet sources. However, it is essential to acknowledge the limitations associated with access to GPT-4, as it necessitates a membership subscription (\$20 per month), which may restrict its availability for some educators and institutions. While the findings indicated that the tool improved the efficiency of feedback provision, the authors emphasized the crucial role of teacher involvement in this process, as it is imperative for teachers to consistently evaluate and refine the generated feedback before sharing it with their students.

Moreover, the authors highlight the importance of crafting specific and well-defined prompts when utilizing LLMs, as the quality of the outputs can be greatly influenced by the input provided. This underscores the value of educators being actively involved in the feedback process to ensure that the response generated by the LLMs meets educational standards. The importance of formulating intentional and detailed prompts, identified as an essential component in generating high-quality outputs, has also been acknowledged in various studies (e.g., Du & Hashimoto, 2025; Kohnke et al., 2023). Thus, the integration of LLMs in language learning contexts must be approached with a balanced perspective that acknowledges both their potential benefits and the essential role of teachers in ensuring their accuracy and usefulness.

In terms of materials development using LLMs, Shin and Lee (2023) investigated whether ChatGPT could generate language assessment materials as effectively as those created by human teachers. The authors emphasized that LLMs can assist second language (L2) teachers with labor-intensive tasks, such as creating assessment items, by emphasizing their ability to generate and design unique texts. To assess LLMs' usefulness for L2 teachers, participants (pre- and in-service L2 teachers) in this study evaluated language test items created by humans and ChatGPT through a blind test. The findings revealed that ChatGPT generated test items with a natural flow and expressions comparable to those created by humans. However, human-created test items were considered more attractive than those generated by ChatGPT, as human experts created more plausible distractors (i.e., more challenging), which heightened test-taker engagement.

Based on these results, the authors argue that ChatGPT can generate natural assessment materials, potentially reducing teachers' workload, yet teacher involvement remained essential for revising the items. Although previous research on using ChatGPT in developing teaching materials showcased positive results (e.g., Carlson et al., 2023; Shin & Lee, 2023), its use remains largely unanchored in a pedagogical framework. Therefore, AILL studies that provide insights into teaching materials development grounded in strong pedagogical principles and research foundations are essential.

TESOL'S 6 PRINCIPLES AND THE APPLICATION OF AI IN LANGUAGE LEARNING

While developed for English language instruction, the TESOL International Association's 6 Principles (Hellman et al., 2019) provide a broad framework for effective language teaching for both general and specialized learners (e.g., in English for Specific Purposes), rooted in decades of research on language pedagogy and acquisition. We focus on three core principles: Principle 1: Know your learners, Principle 2: Create conditions for language learning, and Principle 3: Design high-quality lessons for language development, as shown in Table 1. Hellman et al. (2019) argue that these principles "carry substantial weight because their effective implementation can make a significant difference for learner success" (p. 7). These principles are founded on research-based strategies that address learners' individual needs, enhance learning environments, and structure lessons to support ongoing language development. Aligning prompts with these principles ensures that assessment items are supported by proven pedagogical practices.

Table 1

TESOL International's Principles and genAI Assessment Applications

TESOL Principle (Hellman et al., 2019)	Key Strategies	GenAI Application in Assessment
Principle 1: Know Your Learners	Understand learners' needs; integrate their backgrounds to make learning more meaningful and engaging.	Incorporate learners' life experiences into prompts and create culturally relevant assessment items.
Principle 2: Create Conditions for Language Learning	Define and explain expectations to reduce foreign language anxiety; promote positive washback.	Generate frequent low-stakes formative assessments that align with clear learning objectives.
Principle 3: Design High-Quality Lessons for Language Development	Build on existing knowledge and skills; define what to assess and why.	Incorporate cultural references relevant to learners; select appropriate test item types.

The first principle, "Know your learners," emphasizes understanding learners' needs and backgrounds to facilitate active engagement. It stresses the importance of integrating students' L1 and cultural background into lessons to promote an engaging and inclusive learning environment where learners can take advantage of their existing cultural knowledge. Hellman et al. (2019) recommend using needs analyses, a process for identifying learners' specific goals and language requirements, to better understand learners and the types of English they may encounter.

As research shows, learners' needs vary greatly depending on the discourse communities they aim to join, from academic settings to workplaces (Long, 2005). Recognizing learners' needs and backgrounds is vital in designing equitable and context-sensitive assessments. GenAI tools like ChatGPT can support this process by enabling educators to quickly create language assessments that adapt to students' backgrounds and the

cultural contexts in which they are learning English. This also agrees with the language assessment principle that “test content should reflect sensitivity to candidates’ socio-cultural norms” (Hughes & Hughes, 2020, p. 37). By embedding details from a needs analysis directly into prompts, educators can harness genAI to create equitable assessments that respect learners’ diverse experiences.

The second principle, “Create conditions for language learning,” encourages teachers to be aware of learners’ foreign language anxiety and to create a learning environment in which learners feel secure and safe to take risks freely. To reduce learner anxiety, teachers can clearly communicate expectations for success, which in assessment means sharing specific learning objectives and desired outcomes. Moreover, frequent low-stakes tests can familiarize students with specific question types, reducing the anxiety they might experience during high-stakes tests. genAI tools can streamline the creation of customized assessments, making it easier for educators to incorporate frequent low-stakes tests without the significant time investment typically required.

In the third principle, “Design high-quality lessons for language development,” the high-quality lesson is defined as being built on “learners’ existing knowledge and skills, stimulating internal cognitive-linguistic processes through learners’ active engagement” (Hellman et al., 2019, p. 40). In line with this principle, genAI can enhance lesson quality by rapidly creating assessments that incorporate cultural references, settings, themes, or realistic situations. When teachers acknowledge learners’ educational and occupational histories, life goals, motivations, and interests, they can align instruction with students’ existing knowledge and future aspirations, fostering a more meaningful and engaging learning experience.

Educators can bring these principles into practice by structuring their prompts to reflect each one directly. For example, Principle 1, “Know Your Learners,” can be addressed by including specific details about students’ backgrounds and cultural contexts in the prompt. Principle 2, “Create Conditions for Language Learning,” can be put into practice by specifying the low-stakes, formative nature of the assessment and its clear objectives. Finally, Principle 3, “Design High-Quality Lessons,” is enacted in defining the target skills, lesson topics, and the specific test item types we want the genAI to use. Building a prompt this way elevates it from a simple instruction into a true partnership with the technology, helping us create more meaningful and pedagogically sound materials.

Applying these principles effectively requires teachers to utilize the technology with clear, well-structured prompts based on sound testing pedagogy. According to Hughes and Hughes (2020), a valid test “measures accurately what it is intended to measure” (p. 29). The quality of assessments generated by genAI relies on the clarity and precision of the prompts provided. Poorly designed or vague prompts will produce low-quality test items. An essential first step in testing is clearly defining what one wants to assess and why.

The types of test items (e.g., multiple choice, short answer, true/false, gap-fill, etc.) should elicit the behaviors and skills they aim to measure. For example, a direct assessment of writing ability involves writing tasks, such as short-answer questions or essays. Conversely, multiple-choice questions are an indirect assessment, as they tend to measure students’ receptive and recognition knowledge rather than productive skill. This careful selection of test items also promotes positive washback by encouraging practice of the skills being tested, but creating an effective assessment ultimately depends on balancing these choices with factors like test validity, reliability, practicality, and their impact on teaching and learning (Hughes & Hughes, 2020).

USING GENAI TO DEVELOP LANGUAGE ASSESSMENTS

To put these principles into practice, it is important to remember that genAI is a valuable tool for developing language assessments, provided its use is guided by well-defined principles, clear objectives, and adherence to pedagogical standards. A principle-driven approach ensures assessments are accurate, effective, and aligned with educational best practices. Frequent, low-stakes assessment is an effective teaching strategy across various educational levels and settings (e.g., Black & William, 1998; Colby-Kelly & Turner, 2008). Tests do more than just measure knowledge; they also act as a learning event. The repeated practice of testing and retrieving knowledge aids students in retaining information over the long term (Kirschner et al., 2022). Formative assessments enable students to evaluate their own learning and help instructors identify students facing difficulties. Assessments, in turn, allow instructors to adjust instruction accordingly. Frequent low-stakes testing is also crucial in environments where students face numerous high-stakes English exams. Low-stakes tests enhance comprehension and activate the mental processes needed for success in high-stakes exams (Adesope et al., 2017).

Creating effective low-stakes and formative assessments in the language classroom requires careful alignment of learning outcomes with the assessments themselves. To do this, instructors must identify the specific skills they intend to evaluate and then design tests that clearly demonstrate students' understanding of those skills. Frequent formative assessments are valuable, but creating short quizzes, multiple-choice tests, or short-answer questions is challenging and time-consuming. While more institutional support is the ideal solution for teacher workload, genAI offers a practical way to assist. It automates the initial drafting of assessments, freeing up educators to apply their professional judgment where it is irreplaceable: namely, refining lesson content, analyzing student performance, and providing personalized feedback. To help alleviate the burden of material creation, genAI can support instructors in developing practice tests, quizzes, and short-answer exams.

To translate these concepts of principle-driven, low-stakes assessment into practice, the following prompt template offers a direct, copy-and-paste tool for educators. It is designed to be a clean and practical starting point built on the core pedagogical framework. The template directly integrates Principle 1 (Know Your Learners) by prompting for specific details about student backgrounds and experiences; it reflects Principle 2 (Create Conditions for Language Learning) by framing the task as a low-stakes "formative assessment" with clear outcomes; and it follows Principle 3 (Design High-Quality Lessons) by requiring the educator to define the lesson content and specific assessment techniques.

Example prompt template:

1. As an expert language test developer, you are tasked with constructing a formative assessment for learners of **[target language]** at **[proficiency levels]**. These learners are **[age range]** years old and come from **[cultural backgrounds]**.
2. The quiz will cover **[describe your topic]**. Include topics from previous lessons such as **[topics]**. The questions should be written in authentic language that the learners would encounter in **[learners' context]**. The ultimate goal for these students is to be able to **[clear learning outcomes]**.
3. To make the test both meaningful and relevant, incorporate elements from the learners' cultures, educational backgrounds, and life experiences into the questions or scenarios. **[Relevant background information]**.

4. *[When creating the multiple-choice questions, ensure that the alternate responses are plausible and competitive, pushing the learners to think critically. Avoid using an “all of the above” option.] Include the answer key at the end.*

Implementing genAI in Language Assessments for a University ESP Program

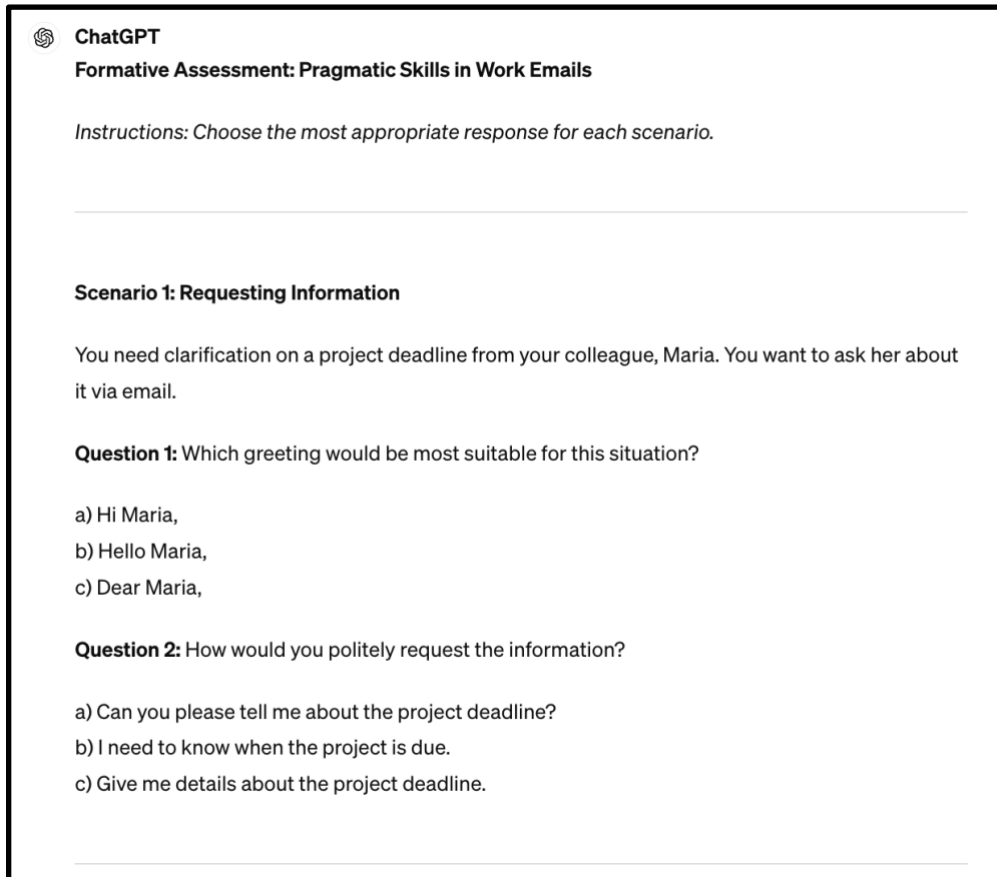
This section will demonstrate how TESOL International’s principles for exemplary teaching (Hellman et al., 2019) can guide an instructor’s use of genAI to create formative language assessments within a university English for Specific Purposes (ESP) program. For context, this program provides weekly English classes for university employees, mainly custodians. ESP learners are usually familiar with English and are learning the language to communicate professional skills and perform profession-related activities (Rahman, 2015). Most of these students are in their 50s-60s and first-language (L1) Spanish speakers from various Latin American countries. The course goals are based on these students’ work needs; the class topics include but are not limited to English for emails, safety, and giving directions relevant to the university.

To illustrate the process for readers, the instructor developed a quiz on email pragmatics using the previously described prompt template and ChatGPT-4. This lesson’s objective is for the students to be able to email their supervisor requesting a day off. Background information about the students was collected from a needs analysis. For instance, many students enjoy traveling to visit family back home, and more than half of them struggle with using English in medical contexts. This information was included in the prompt to make the assessment more relevant to the students.

Application with ChatGPT

Using the prompt template and context from the ESP example case, ChatGPT-4 generated an eight-question quiz with an answer key, as shown in Figure 1. The quiz covered four scenarios illustrating English use in emails. This assessment, or specific parts of it, can be integrated into the lesson in various ways, either as a review at the beginning of class or as an exit ticket to gauge students’ understanding and to identify areas where they may need further support.

Figure 1

Initial Quiz Produced Using Example Prompt Template

ChatGPT
Formative Assessment: Pragmatic Skills in Work Emails

Instructions: Choose the most appropriate response for each scenario.

Scenario 1: Requesting Information

You need clarification on a project deadline from your colleague, Maria. You want to ask her about it via email.

Question 1: Which greeting would be most suitable for this situation?

- a) Hi Maria,
- b) Hello Maria,
- c) Dear Maria,

Question 2: How would you politely request the information?

- a) Can you please tell me about the project deadline?
- b) I need to know when the project is due.
- c) Give me details about the project deadline.

This initial output is a starting point from which many improvements can be made. For example, a teacher reviewing Question 1 might recognize that without additional context on the workplace culture or the sender's relationship with their colleague, all the greetings could be appropriate. Such refinement also involves ensuring the assessment reflects learners' cultural backgrounds and professional demands. The original quiz lacked questions about medical contexts, which were relevant to the students' personal priorities as identified by the beginning of the semester's needs analysis. One of the questions in the needs analysis was, "Explain a situation where you really needed to communicate something in English but could not." More than half of the students mentioned struggling with knowing what to say at the doctor's office. Changing assessment topics is just one way to make adjustments and improvements based on student and classroom needs. In this case, after reviewing the initial output, the genAI was prompted to "create scenarios and questions in a medical context." It then generated eight additional questions focused on medical-related email communications, illustrated in Figure 2.

Figure 2

Follow-Up Interaction Adjustment to Meet Student Needs

Scenario 1: Making an Appointment

You need to schedule a doctor's appointment for a routine check-up. You want to email the clinic to book an appointment.

Question 1: Which greeting would be most suitable for this situation?

a) Hi,
b) Hello,
c) Dear Doctor,

Question 2: How would you request the appointment in a polite manner?


a) I need to see the doctor.
b) Can I have an appointment?
c) May I please schedule a doctor's appointment for a routine check-up?

Scenario 2: Requesting Prescription Refill

You need a refill for your prescription medication. You decide to email your healthcare provider to request it.

Question 3: Which opening would be most appropriate for this email?


a) I need more medicine.
b) Can you refill my prescription?
c) I hope you're doing well.



In many learning contexts, students have varying levels of proficiency. In this ESP example case, some students have been using English professionally for years, and others have only recently begun learning. GenAI can assist teachers in implementing differentiated instruction, supporting Principle 2, “creating conditions for language learning.” To do this, ChatGPT was prompted to edit the assessment: “Some of these students just started learning English. Translate the questions and instructions into Spanish but leave the answer choices in English,” as demonstrated in Figure 3. This is a quick way to produce differentiated but similar assessments that are customized to meet the requirements of individual students.

Figure 3

Translated Version of the Quiz for Differentiated Learning

 **ChatGPT**
Evaluación Formativa: Habilidades Pragmáticas en Correos Electrónicos Laborales

Instrucciones: Elija la respuesta más adecuada para cada escenario.

Escenario 1: Solicitar Información

Necesita aclarar la fecha límite de un proyecto con su colega, María. Quiere preguntarle sobre esto por correo electrónico.

Pregunta 1: ¿Qué saludo sería más adecuado para esta situación?


a) Hi Maria,
b) Hello Maria,
c) Dear Maria,

Pregunta 2: ¿Cómo solicitaría la información de manera educada?

a) Can you please tell me about the project deadline?
b) I need to know when the project is due.
c) Give me details about the project deadline.

Escenario 2: Confirmar una Reunión

Necesita confirmar una hora de reunión con su supervisor, el Dr. Rodríguez, para la próxima semana.



This example offers a small glimpse of how genAI, guided by established principles, can support tailored and effective language assessments while adapting to diverse learner needs. The ultimate success of these technological tools lies in their thoughtful integration by knowledgeable educators. While genAI offers immense potential, the educators' guidance and evaluation ensure technology serves as a complement to, rather than a replacement for, principled pedagogical planning.

PEDAGOGICAL OUTCOMES AND REFLECTIONS

In practice, the main benefit of employing this approach was not merely time efficiency, but also the ability to quickly adapt materials based on what was happening in the classroom. For example, after administering a quiz, the instructor noticed that a question had multiple correct answers and that some students struggled with the English instructions. By utilizing genAI, it was possible to immediately create a revised, more accurate version and translated copies for

the next class. This ongoing process of creation, evaluation, and enhancement facilitated responsiveness that is difficult to achieve with conventional materials preparation.

Additionally, this process highlighted that while utilizing genAI was faster, it was not necessarily easier, as it required constant critical evaluation from the instructor. The initial output from the genAI was rarely ready to use as is; instead, it served as a starting point that required a teacher's evaluation and modification. For instance, a first draft of a genAI-generated quiz was enhanced by incorporating specific details about a local hospital that students would easily recognize, thus making the assessment more relevant. Based on this instructor's experience and observations, genAI's greatest strength may be its ability to support an educator's capacity to be reflective and highly responsive to their students' needs.

CONCLUSION

This paper explored how emerging genAI technologies, specifically ChatGPT, can be integrated into language teaching to support the development of formative language assessments. As previous studies suggest (Carlson et al., 2023; Shin & Lee, 2023), genAI offers benefits in generating personalized feedback and labor-intensive material development. However, limitations resulting in less engaging output (Shin & Lee, 2023) and outputs with false information highlight the essential role of teachers. When approached with thoughtful guidelines, such as TESOL International's 6 Principles (Hellman et al., 2019), genAI tools can enhance educators' capacity to create meaningful assessment opportunities.

To facilitate this integration, numerous free resources and training materials, such as aiforeducation.io and Google's "Generative AI for Educators" course, are available to guide teachers in refining their application of these technologies. Nevertheless, future research is needed to systematically evaluate the effectiveness of these tools, as well as the quality of associated training programs, to ensure teachers employ genAI solutions effectively and responsibly. Such work should also address emerging ethical issues like data confidentiality, algorithmic bias, the environmental impact of these tools, and challenges related to human labor. Regarding human labor, it is crucial to address concerns that organizations may use genAI to replace rather than support educators. As Conrad (2024) suggests, proactively involving teachers in creating AI policies and providing meaningful professional development can ensure that technology augments, rather than automates, their work. Moreover, empirically examining ChatGPT's direct impact on students' language development will provide a stronger evidence base for its inclusion or exclusion in teaching practices.

Finally, by demonstrating the use of carefully designed prompts within a specific instructional context, we have illustrated the positive potential of genAI technologies to create low-stakes, formative assessments. While the aforementioned genAI-related issues always need to be considered, we believe that principled use of genAI technologies has the potential to streamline material development, promote language learning, and reduce teachers' workloads.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Dr. Matt Kessler for his valuable feedback and encouragement, which helped shape this paper.

REFERENCES

- Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, 87(3), 659-701.
<https://doi.org/10.3102/0034654316689306>
- Agyapong, B., Obuobi-Donkor, G., Burbach, L., & Wei, Y. (2022). Stress, burnout, anxiety and depression among teachers: A scoping review. *International Journal of Environmental Research and Public Health*, 19(17), 10706. <https://doi.org/10.3390/ijerph191710706>
- Ahn, Y. Y. (2023). Performance of ChatGPT 3.5 on CSAT: Its potential as a language learning and assessment tool. *Journal of the Korean English Education Society*, 22(2), 119–145.
- An, X., Chai, C. S., Li, Y., Zhou, Y., & Yang, B. (2023). Modeling students' perceptions of artificial intelligence assisted language learning. *Computer Assisted Language Learning*, 1-22.
<https://doi.org/10.1080/09588221.2023.2246519>
- Berg, G., & Plessis, E. (2023). ChatGPT and generative AI: Possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. *Education Sciences*, 13(10), 998.
<https://doi.org/10.3390/educsci13100998>
- Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148.
- Carlson, M., Pack, A., & Escalante, J. (2023). Utilizing OpenAI's GPT-4 for written feedback. *TESOL Journal*, e759. <https://doi.org/10.1002/tesj.759>
- Carter, S., & Nielsen, M. (2017). Using artificial intelligence to augment human intelligence. *Distill*, 2(12), <https://doi.org/10.23915/distill.00009>
- Chen, Y. C. (2022). Effects of technology-enhanced language learning on reducing EFL learners' public speaking anxiety. *Computer Assisted Language Learning*, 1–25.
<https://doi.org/10.1080/09588221.2022.2055083>
- Colby-Kelly, C., & Turner, C. E. (2008). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *The Canadian Modern Language Review*, 64(1), <https://doi.org/10.3138/cmlr.64.1.009>
- Dillenbourg, P. (2016). The evolution of research on digital education. *International Journal of Artificial Intelligence in Education*, 26(2), 544–560. <https://doi.org/10.1007/s40593-016-0106-z>
- Du, Z., & Hashimoto, K. (2025). Exploring sentence-level revision capabilities of large language models in English for academic purposes writing assistance. *Asian-Pacific Journal of Second and Foreign Language Education*, 10(27), <https://doi.org/10.1186/s40862-025-00334-z>
- Hellman, A. B., Harris, K., & Wilbur, A. (2019). *The 6 principles for exemplary teaching of English learners®: Adult education and workforce development*. TESOL Press.
- Higton, J., Leonardi, S., Choudhury, A., Richards, N., Owen, D., & Sofroniou, N. (2017). *Teacher workload survey 2016*. Department for Education.
- Holstein, K., McLaren, B. M., & Alevin, V. (2019). Co-Designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics*, 6(2), 27–52.
<https://doi.org/10.18608/jla.2019.62.3>
- Holstein, K., & Alevin, V. (2022). Designing for human-AI complementarity in K-12 education. *AI Magazine*, 43(2), 239-248. <https://doi.org/10.1002/aaai.12058>
- Hrastinski, S., Olofsson, A. D., Arkenback, C., Ekström, S., Ericsson, E., Fransson, G., Jaldemark, J., Ryberg, T., Öberg, L., Fuentes, A., Gustafsson, U., Humble, N., Mozelius, P., Sundgren, M. & Utterberg, M. (2019). Critical imaginaries and reflections on artificial intelligence and robots in post digital K-12 education. *Post digital Science and Education*, 1(2), 427- 445. <https://doi.org/10.1007/s42438-019-00046-x>
- Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237–257. <https://doi.org/10.1111/jcal.12610>
- Hughes, A. Hughes, J. (2020). *Testing for language teachers* (3rd ed.). Cambridge University Press.
- Ji, H., Han, I., & Ko, Y. (2023). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, 55(1), 48-63.
<https://doi.org/10.1080/15391523.2022.2142873>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023).

- ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kirschner, P., Hendrick, C., & Heal, J. (2022). *How teaching happens: Seminal works in teaching and teacher effectiveness and what they mean in practice*. Routledge.
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *Relc Journal*, 54(2), 537-550. <https://doi.org/10.1177/00336882231162868>
- Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability*, 15(7), 5614. <https://doi.org/10.3390/su15075614>
- Liu, G., & Ma, C. (2024). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching*, 18(2), 125-138. <https://doi.org/10.1080/17501229.2023.2240316>
- Long, M. (2005). *Second Language Needs Analysis*. Cambridge University Press.
- Minnillo, S., Jones, L., & Garcia, S. (2024). AI in the L2 classroom: Serving language educators through professional development. *L2 Journal*, 16(1), 1-14. <https://doi.org/10.5070/L2.21198>
- Moorhouse, B. L., & Kohnke, L. (2024). The effects of generative AI on initial language teacher education: The perceptions of teacher educators. *System*, 122, 103290. <https://doi.org/10.1016/j.system.2024.103290>
- Ngo, T. N., & Hastie, D. (2025). Artificial intelligence for academic purposes (AIAP): Integrating AI literacy into an EAP module. *English for Specific Purposes*, 77, 20-38. <https://doi.org/10.1016/j.esp.2024.09.001>
- Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*, 40(2), 269-275. <https://doi.org/10.1016/j.ijresmar.2023.03.001>
- Qin, F., Li, K., & Yan, J. (2020). Understanding user trust in artificial intelligence-based educational systems: Evidence from China. *British Journal of Educational Technology*, 51(5), 1693-1710. <https://doi.org/10.1111/bjet.12994>
- Rahman, M. (2015). English for specific purposes (ESP): A holistic review. *Universal Journal of Educational Research*, 3(1), 24-31.
- Shin, D. (2023). A case study on English test item development training for secondary school teachers using AI tools: Focusing on ChatGPT. *Language Research*, 59(1), 21-42. <https://doi.org/10.30961/lr.2023.59.1.21>
- Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts?. *Language Learning & Technology*, 27(3), 27-40. <https://www.lltjournal.org/item/1086/>
- Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching*, 6(1), 31-40. <https://doi.org/10.37074/jalt.2023.6.1.17>
- Wang, X., & Reynolds, B. L. (2024). Beyond the books: Exploring factors shaping Chinese English learners' engagement with large language models for vocabulary learning. *Education Sciences*, 14, 496. <https://doi.org/10.1186/s40862-025-00334-z>
- Weng, X., & Chiu, T. K. (2023). Instructional design and learning outcomes of intelligent computer assisted language learning: Systematic review in the field. *Computers and Education: Artificial Intelligence*, 4, 100117. <https://doi.org/10.1016/j.caeai.2022.100117>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 1-27. <https://doi.org/10.1186/s41239-019-0171-0>