Educational Innovation Research, 2026, Volume 4, Issue 7
https://bbwpublisher.com/journal/EIR
Online ISSN: 3029-1852
Print ISSN: 3029-1844

Assessing Spanish Grammar Learning in the Age of Generative AI: A Conceptual Framework

Kejian Qian, Yiduo Wang*, Zhining Wei

School of Foreign Languages, Jilin Normal University, Siping 136000, Jilin, China

*Corresponding author: Yiduo Wang, 361510464@qq.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: This study proposes an approach to assessing foreign language grammar learning when generative artificial intelligence (GenAI) is used during instruction. Focusing on university-level Spanish grammar courses in China, it uses the form–meaning–use model of grammar to organize assessment into three stages: an independent baseline task, AI-assisted practice, and an independent transfer task. A teaching unit on the contrast between the preterite (*pretérito indefinido*) and the imperfect (*pretérito imperfecto*) illustrates how this approach can be used to assess students' responses to AI feedback and compare their AI-assisted work with their performance on an unfamiliar task completed without AI assistance. This comparison helps teachers determine what students can do with AI support and what they can do independently, providing a clearer basis for subsequent instruction.

Keywords: Generative artificial intelligence; Grammar learning assessment; Learning outcomes; Spanish as a foreign language

Online publication: July 26, 2026

1. Introduction

Generative artificial intelligence (GenAI), particularly in the form of large language models (LLMs), is changing how students complete foreign language learning tasks. In grammar learning, students increasingly use these tools to understand grammatical rules and improve their answers. However, an accurate answer produced with AI assistance does not necessarily show that the learner can use the relevant grammar independently. This creates a problem for traditional assessment based mainly on students' final answers: teachers cannot readily determine whether a correct response reflects the learner's own grammatical knowledge or the support provided by AI.

This issue is particularly prominent in the teaching of Spanish grammar, as learners must not only select the correct forms but also apply them appropriately within specific contexts. To address this challenge, this study deconstructs the assessment dimensions for university-level Spanish grammar courses, drawing on Larsen and

Freeman's [1] three-dimensional grammar framework of form–meaning–use. The study establishes an evidence-based assessment framework that integrates independent performance with human-computer collaborative processes, and illustrates its operational trajectory through a teaching case study comparing the Spanish simple past and imperfect tenses, with the aim of providing a theoretical basis and practical guidance for assessment in AI-integrated educational contexts. The contribution of this paper lies in linking the form–meaning–use dimensions of grammar learning to the specific conditions under which task performance occurs. The proposed framework clarifies the evidence that each task can provide, as well as how this evidence should inform teaching decisions.

2. Literature review

Studies have examined the effects of artificial intelligence (AI) on foreign language learning outcomes. Wang et al. (2026) [2] conducted a meta-analysis of empirical studies to estimate the overall effect of AI on foreign language learning outcomes. Focusing on generative AI (GenAI), Pan et al. (2026) [3] reviewed 32 empirical studies in foreign language education. The studies were concentrated in higher education, and writing support was the most common application.

In addition to application contexts, some scholars have focused particularly on the operational definition and psychometric measurement of learning outcomes. In a meta-analysis of 63 studies in higher education, Luo et al. (2025) [4] deconstructed learning outcomes into three dimensions: cognitive, skills-based and affective. They noted that whilst AI tools have produced relatively consistent positive effects on the acquisition of declarative knowledge and specific affective outcomes, research findings regarding deep cognitive processing and the development of higher-order skills remain inconclusive. As emphasised by Weng et al. (2024) [5], the deep integration of GenAI has prompted higher education to re-examine existing assessment models and the core learning constructs they target. In the field of foreign language assessment, Huang et al. (2026) [6] revealed a statistically significant correlation between students' acceptance of GenAI and their standardised test scores (College English Test Band 4). However, this statistical correlation merely indicates that willingness to adopt technology and academic achievement represent two independently measurable constructs; it does not provide direct theoretical or practical support for the notion that technology acceptance directly drives improvements in language proficiency.

Other scholars have examined the validity of inferences drawn from learning evidence when students use AI to complete tasks. Memarian and Doleck (2024) [7] emphasise that collecting data throughout the learning process does not, by itself, establish that learning has occurred. In academic writing, Kolade et al. (2024) [8] found that ChatGPT-generated essays could receive high marks. This finding raises the question of how far the quality of a submitted text reflects the student's own knowledge and ability. Although findings from academic writing cannot be directly generalised to grammar learning, they raise a similar question: whether an AI-assisted response provides evidence of the learner's independent ability. In summary, studies of AI-assisted learning have examined test scores, skill development and learner attitudes. These measures address different aspects of learning and support different interpretations. Therefore, course-level assessment needs to distinguish performance with AI assistance from what learners can do independently and specify which conclusions the available evidence supports.

3. Conceptualising learning outcomes: A three-dimensional grammar perspective

Learning outcomes in university-level Spanish grammar should not be judged solely by students' recall of explicit rules or their accuracy in controlled exercises. Larsen-Freeman's (2017) [1] framework describes

grammar in terms of three interrelated dimensions: form, meaning and use. Form concerns the structure and accurate formation of grammatical expressions. Meaning concerns the meanings those forms convey. Use concerns whether a grammatical choice is appropriate to the communicative context, including when and why a particular form is used [1]. Consequently, although the three dimensions are closely related, accuracy alone may not show whether students understand a grammatical distinction or can use it appropriately in context.

On this basis, the present study uses the following three dimensions to assess grammar learning:

(1) Recognition and production of grammatical forms: assessing whether learners can accurately recognise and produce the target morphosyntactic forms.
(2) Understanding and interpretation of grammatical meaning: assessing whether learners can explain semantic differences between contrasting structures rather than merely repeat memorised rules.
(3) Appropriate use in context: assessing whether learners' grammatical choices fit their communicative intentions and the discourse context.

Although the three dimensions are related, performance in one does not establish mastery of the others. Accurate forms may result from applying memorised rules, while an accurate explanation of meaning does not guarantee appropriate use in a new production task.

Larsen-Freeman's discussion of "inert knowledge" provides a further reason to assess use alongside form and meaning. She notes that learners may have difficulty applying rules in spontaneous communication even when they appear to have mastered those rules in controlled exercises [1]. She therefore presents "grammaring" as an ability developed through practice, rather than treating grammar solely as declarative knowledge. Accordingly, assessment in university-level Spanish grammar courses should examine whether learners can apply their knowledge of form and meaning in contextualised language production.

4. A Framework for assessing grammar learning with GenAI support

4.1. Learning objectives, task performance and task conditions

The framework uses the dimensions of form, meaning and use to guide the assessment of students' grammatical performance. In assessment practice, teachers should design tasks around clear learning objectives to examine whether students can produce the target forms accurately, explain their meanings and choose forms appropriate to the context. Although a single integrated task may provide evidence across all three dimensions, teachers should assess performance in each dimension separately and address the findings in their formative feedback. If a composite score is used, it should be accompanied by comments on students' strengths and difficulties in each dimension.

In addition, interpreting students' performance requires attention to the conditions under which they complete a task, including whether they use GenAI. Students may produce the same answer independently, revise it in response to AI feedback or copy an AI-generated response. However, these different ways of completing the task provide different evidence of what students can do independently. Bachman and Clark (1987) [9] note that language test performance is influenced by both learners' communicative competence and test method factors. Accordingly, the framework records students' performance in form, meaning and use alongside whether they used AI, what assistance they received and how they used that assistance.

4.2. Evidence collection across three tasks

To compare students' performance before, during and after AI assistance, the framework uses three complementary tasks:

(1) Independent baseline diagnostic task: Students complete exercises on the target grammatical structures without AI assistance. Teachers record which forms students produce accurately, which differences in meaning they can explain and where their grammatical choices are inappropriate to the context. These responses provide a baseline for comparison with the later tasks.

(2) AI-assisted learning task: Teachers collect students' initial responses, prompts, AI feedback, revised responses and written explanations of their revisions. These records allow teachers to examine whether students evaluate suggestions against the task context, identify incorrect grammatical explanations or inappropriate revisions, and explain why they accept or reject the feedback. Teachers also check the accuracy of the AI feedback and compare it with students' explanations to identify responses that merely repeat the tool's wording without explaining the grammatical choice.

(3) Independent transfer task: Following AI-assisted learning, students complete a new task without AI assistance. The task targets the same grammatical features but uses different language materials in a different communicative context. Teachers then assess performance in form, meaning and use to examine whether students can apply what they have practised without AI support. The task should require students to apply the grammatical distinction in a new context, rather than reproduce a previous answer with different vocabulary. In addition, its difficulty should be comparable to that of the baseline task so that differences in performance are less likely to reflect task difficulty. Where the teaching schedule allows, a delayed post-test can also assess retention over time.

For each task, the teacher records four items: (1) whether the learner produces the target form accurately; (2) whether the learner explains the meaning conveyed by that form in the given sentence; (3) whether the choice fits the wider discourse context and the learner's intended meaning; and (4) whether the learner uses AI and, if so, what assistance they receive and how they use it. The first three items describe performance in form, meaning and use, while the fourth documents the AI assistance involved in completing the task.

4.3. Interpreting Evidence across Tasks

Comparing students' initial responses, AI-assisted revisions and independent transfer performance helps teachers identify difficulties that persist across tasks and those that emerge when AI support is removed. Drawing on Kane's (2013) [10] argument-based approach to validation, the framework asks teachers to identify the evidence supporting an interpretation and consider other explanations for the observed performance. For example, a student may accept an AI correction to a grammatical form without understanding its meaning in the sentence. If the student cannot explain the change, the teacher can ask them to compare the original and revised forms in context to identify what remains unclear.

Comparing performance across the three tasks can help teachers identify areas for further instruction:

(1) If students can explain grammatical rules accurately but repeatedly make errors in new communicative contexts, subsequent instruction should focus on applying those rules. Activities can combine comparisons of grammatical choices in different contexts with contextualised production.

(2) If students provide accurate explanations with AI assistance but make errors in the independent transfer task, teachers should ask them to explain their choices in further tasks without AI. Comparing these

explanations with their responses can help teachers decide whether to revisit the meaning of the forms or provide more practice in contextual use.

(3) If students perform well in form, meaning and use across all three tasks, teachers can introduce more varied communicative contexts to extend their use of the target structures.

4.4. Using assessment evidence to inform teaching

In formative assessment, teachers use evidence from students' responses to plan subsequent instruction. They can group recurring difficulties according to the three dimensions. Errors in grammatical form call for focused practice in producing the target forms; difficulty explaining grammatical meaning calls for comparisons of the meanings conveyed by contrasting forms; and problems with use call for tasks in which students choose and use forms in different communicative contexts. Furthermore, if AI feedback contains recurring errors or misleading explanations, teachers should correct these in class, revise the guidance on prompting and show students how to check suggestions against grammatical rules and the task context.

As Black and Wiliam (2009) [11] emphasise, evidence of learning has formative value when teachers use it to inform subsequent instruction. Accordingly, teachers specify the task conditions, collect evidence of performance in form, meaning and use, and interpret that evidence to adjust subsequent tasks. They then examine students' responses to the revised tasks to decide whether further changes to instruction are needed.

5. Applying the framework: An illustrative Spanish grammar lesson

To illustrate how the framework can be applied, this section presents a 45-minute lesson on the Spanish preterite–imperfect contrast, followed by a delayed task one week later. The lesson asks students to form the two tenses accurately and explain how their choices affect the interpretation of events in a narrative. The teacher prepares narrative materials for two different contexts and asks students to keep their initial independent responses, records of their interactions with the GenAI tool, revised responses and explanations of their revisions. These materials provide a basis for comparing performance across the three tasks.

(1) Phase 1: Independent baseline diagnosis

Without AI assistance, students complete a context-based gap-fill task and explain their choice in writing:

Ayer, cuando entré en la biblioteca, Ana ___ (leer) una carta. No levantó la vista hasta que me senté a su lado.

After collecting students' independent responses, the teacher asks them to consider the narrative perspective: was Ana already reading when the narrator entered the library, or is her reading presented as a subsequent event? The teacher then replaces *leía* with *leyó* and asks students to compare how the two forms present the event. Students also consider whether the presence of ayer is sufficient to determine the tense. The responses collected before this discussion provide the baseline evidence of students' performance in grammatical form, aspectual meaning and narrative use.

(2) Phase 2: AI-assisted learning

In this phase, students enter the original sentence, the alternative version and their initial explanation into a GenAI tool. They ask it to explain how *leía* and *leyó* present the events differently and whether each form fits the surrounding narrative. On a worksheet, students record the following: (a) the AI's suggestions; (b) textual evidence supporting or challenging them; and (c) their reasons for accepting or rejecting the feedback.

The teacher provides questions to guide students' evaluation of the feedback. For example, does the AI treat a temporal adverbial as requiring a particular tense, or claim that the imperfect means the action was never completed? Following a whole-class discussion, students submit their revised responses and explanations of the changes they have made.

(3) Phase 3: Independent transfer and delayed retention assessment

In this phase, students complete the following transfer task without AI assistance:

Cuando llegué a casa, mi hermano ___ (preparar) la cena.

Students complete the sentence and add a continuation for each of two situations: (a) the brother was already preparing dinner when the narrator arrived; and (b) the brother prepared dinner after the narrator arrived. For each version, students explain how their choice of verb form and the continuation establish the relationship between the two events. One week later, the teacher administers a delayed task without AI assistance to assess retention, using a new narrative context that tests the same grammatical contrast.

(4) Phase 4: Analysis of learning evidence

The teacher examines students' initial responses, AI interaction records and subsequent independent responses in terms of form, meaning and use. In the baseline task, producing leía would show that the student can form the required imperfect tense correctly in this item. Explaining that Ana was already reading when the narrator entered would indicate an understanding of the ongoing background activity expressed by the imperfect. For the use dimension, the teacher examines whether the student explains how leía fits the surrounding narrative, including the contextual support provided by *"No levantó la vista hasta que me senté a su lado"*.

During the AI-assisted phase, the teacher also examines how students evaluate feedback, including whether they explain their reasons for accepting a suggestion or identify advice unsupported by the context. The teacher checks whether students explain the grammatical choice rather than simply repeat the AI's wording. Finally, in the independent transfer task, the teacher considers the choice between preparaba and preparó alongside the student's explanation and continuation. An accurate form, an appropriate explanation and a coherent continuation together would support a judgement of successful independent use in that task.

6. Conclusion

When GenAI contributes to a grammar task, the quality of the final answer needs to be interpreted in relation to how the student completed it. This paper therefore proposes a conceptual framework that combines assessment of form, meaning and use with records of AI assistance and subsequent independent performance. Its contribution lies in explaining how teachers can use evidence from different task conditions to judge students' grammar learning and plan further instruction.

The framework links an independent baseline task, an AI-assisted learning task and an independent transfer task. Each provides a different source of evidence: initial responses record students' starting performance; interaction records and revision explanations document how they evaluate and use AI feedback; and transfer responses show how they use the target forms without that assistance. Comparing these materials helps teachers identify whether difficulties concern the production of grammatical forms, the explanation of meaning or appropriate use in context.

The Spanish preterite–imperfect example illustrates how this comparison can inform classroom assessment. Teachers consider students' verb forms alongside their explanations and narrative continuations, then use the resulting evidence to select further practice. Within this framework, records of AI-assisted work

contribute to assessment by allowing teachers to examine the feedback students receive and the reasons they give for their revisions. Subsequent independent responses provide a further basis for deciding which aspects of grammar require additional teaching.

Funding

Jilin Normal University Young Talent Teaching Reform Research Project, "A Study of the Effects of an AI-Enhanced Blended Learning Model on Learning Outcomes among University Foreign Language Majors"; 2024 University-Level Education and Teaching Reform Research Project, "Innovative Pathways for Reforming University Spanish Programs in the Context of Digital Transformation"

Disclosure statement

The author declares no conflict of interest.

References

[1] Larsen-Freeman D, 2017, Just Learning. Language Teaching, 50(3): 425–437.

[2] Wang L, Liu T, Zhao B, 2026, A Meta-Analysis of the Impact of Artificial Intelligence on Foreign Language Learning Outcomes. Foreign Languages Research, 2026(3): 84–91.

[3] Pan H, Li J, Ma X, 2026, Applications and Research of Generative Artificial Intelligence in Foreign Language Teaching: A Systematic Review. Foreign Language Teaching and Research, 58(2): 250–262.

[4] Luo J, Zheng C, Yin J, et al., 2025, Design and Assessment of AI-Based Learning Tools in Higher Education: A Systematic Review. International Journal of Educational Technology in Higher Education, 22: Article 42.

[5] Weng X, Xia Q, Gu M, et al., 2024, Assessment and Learning Outcomes for Generative AI in Higher Education: A Scoping Review on Current Research Status and Trends. Australasian Journal of Educational Technology, 40(6): 37–55.

[6] Huang Z, Jin H, Chen C, 2026, From Acceptance to Achievement: Exploring Foreign Language Learners' Perceptions of Generative AI. Foreign Language Education, 47(1): 52–59.

[7] Memarian B, Doleck T, 2024, A Review of Assessment for Learning with Artificial Intelligence. Computers in Human Behavior: Artificial Humans, 2(1): Article 100040.

[8] Kolade O, Owoseni A, Egbetokun A, 2024, Is AI Changing Learning and Assessment as We Know It? Evidence from a ChatGPT Experiment and a Conceptual Framework. Heliyon, 10(4): e25953.

[9] Bachman LF, Clark JLD, 1987, The Measurement of Foreign/Second Language Proficiency. The ANNALS of the American Academy of Political and Social Science, 490(1): 20–33.

[10] Kane MT, 2013, Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1): 1–73.

[11] Black P, Wiliam D, 2009, Developing the Theory of Formative Assessment. Educational Assessment, Evaluation and Accountability, 21(1): 5–31.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.