appi
ASSOCIAÇÃO PORTUGUESA
DE PROFESSORES DE INGLÊS

eJournal

Spring / Summer 2026
YEAR 9 · 29

ISSN 2184-7525

AI-Aware Assessment for the EFL Classroom

Carlos Lindade

Summary:
Generative AI has not made assessment impossible. It has exposed assessment that was never measuring much. Rather than banning AI tools, the EFL classroom needs assessment that protects validity, that keeps a learner's thinking visible, while building the AI literacy this generation will need. This article starts where the work starts, with classrooms co-constructing responsible-use guidelines, and then sets out three frameworks that fit together: DigComp 3.0, the European reference for digital competence, which now integrates AI and is built to inform assessment; the AILit Framework, its primary and secondary companion; and the AI Assessment Scale, a practical tool for deciding how much AI a task should allow. It closes with ten worked examples for lower and upper secondary English, two for each level of the scale, and a set of design principles.

The real problem is not whether machines think but whether men do. (Skinner, 1969, p. 288)

1. The first step: agreeing on rules... together!

In my previous contribution to this eJournal (Lindade, 2026) I argued for moving from bans to balance, and I want to begin there, because it is also where rethinking assessment in the age of AI begins. The pressing challenge for schools today is not to write regulations that forbid AI. It is to help whole school communities: teachers, students, and families, learn what responsible AI use looks like, and to co-construct guidelines that are meaningful and easy to share.

This matters for assessment more than it first appears. A class that has agreed, together, when AI may be used, when it may not, and why, has already done the groundwork for everything that follows. The guidelines do not need to be elaborate. A simple, co-created agreement, a traffic-light scale that marks tasks green (AI encouraged), amber (AI allowed within limits), or red (no AI), gives students a shared language and gives families a clear document they can understand. Co-construction is the point. When learners help define the boundaries, they internalise them, and a teacher can hold them to account because the rules are collectively owned rather than imposed.

That agreement is the foundation. The rest of this article is about what we put on top of it: tasks that are worth setting in the first place.

2. Who are we teaching, and why does it sharpen the question?

Agreeing how we will use AI together answers one question. It immediately raises a harder one. If these tools are now part of the room, what exactly are we assessing, and can we still see the learner inside the work? To answer that, it helps to be clear about who our learners actually are.

It is tempting to reach for dramatic claims about a generation whose brains have been rewired. One should resist that, because the evidence is thin, and the history of our field is full of confident generational pronouncements that did not survive scrutiny. The "digital natives" idea is the cautionary example: Bennett, Maton, and Kervin (2008) showed that it rested largely on anecdote rather than evidence, and that young people's skills are far less uniform than the label suggests. Any claim about today's learners has to clear a higher bar.

There is, nonetheless, a defensible and specific distinction. Generation Alpha, the students born from 2010 onward (a term coined by the social researcher Mark McCrindle), differ from Generation Z and the millennials in one way that matters for us. Earlier cohorts adopted each technology after their formative years; Generation Alpha are the first to pass through primary and lower secondary schooling while generative AI is simply present, an ordinary feature of the world rather than a novelty to be learned (McCrindle & Fell, 2021). For an English teacher, that is the crucial point. These learners meet a fluent, tireless, language-producing machine before they have built the critical faculties to question it. The discontinuity is not in their neurology. It is in the sequence: exposure arrives before judgement.

And they are not waiting for us. The European Commission and OECD AILit Framework (OECD, 2025) summarises the picture: most young people are already experimenting with AI, in structured and unstructured ways, yet around half of those aged 17 to 27 struggle to identify its shortfalls, such as whether a system can invent facts. The seven-country European survey it draws on, which includes Portugal, found that 74% of 12- to 17-year-olds believe AI will play a significant role in their professional lives, while only 46% feel their school prepares them for it (Vodafone Foundation, 2024). That gap, not a story about attention spans, is the case for acting. Our students are using these tools without guidance, the English classroom is no longer their only source of the language or of

feedback on it, and the question of how we assess them can no longer be postponed.

3. The real question is validity, not cheating

Faced with all this, the instinct many of us reach for first is some version of "how do I stop them?" It is understandable, and it leads nowhere useful. AI detectors are unreliable and produce false accusations; outright bans simply push use out of sight. Both responses spend our energy policing behaviour instead of improving what we ask students to do.

There is a better question, and it predates ChatGPT. Dawson, Bearman, Dollinger and Boud (2024) put it plainly in the title of their paper: validity matters more than cheating. An assessment is valid when it actually measures the learning it claims to measure. If a task can be completed, invisibly, by a machine, the real problem is not that a student was dishonest. The problem is that the task was never measuring much. Generative AI has not broken assessment. It has revealed the assessments that were already weak, and handed us a reason, at last, to redesign them. The rest of this article treats AI not as an enemy to be excluded or a miracle to be embraced, but as something our students live with, and asks how we design English tasks that keep their thinking visible.

4. Three frameworks that fit together

A teacher designing an AI-aware assessment faces three questions, in sequence. What should my learners be able to do with and around AI? What does that look like at their age and stage? And, for this particular task, how much AI should I allow? Each question has a framework built to answer it, and, crucially, the three were designed to interlock rather than compete.

The first question is about **competence**, and the reference point is **DigComp 3.0** (Cosgrove & Cachia, 2025), the fifth edition of the European Digital Competence Framework. This is the European Union's official description of what it means to be digitally competent, organised into five competence areas, twenty-one competences, and four proficiency levels (Basic, Intermediate, Advanced, Highly Advanced). Two features make it the natural backbone for assessment. First, it integrates AI transversally rather than treating it as a separate module, labelling each competence as AI-explicit or AI-implicit, which matches the balanced stance that AI is one digital technology among many. Second, it is built to inform assessment: it is used across Europe to develop assessments and to support both summative and formative assessment through its learning outcomes.

The second question is about **what AI literacy looks like in school**, and here the **AILit Framework** (OECD, 2025) is the companion. Where DigComp describes competence for any citizen across a lifetime, AILit translates AI literacy specifically for primary and secondary learners, across four domains, Engaging, Creating, Managing, and Designing AI, with classroom scenarios for each. For English teachers, the first three domains are the relevant ones; Designing AI belongs more to the computing room. The two frameworks are not rivals but layers of the same project: the Joint Research Centre confirms that DigComp 3.0 was mapped against the AILit competences and covers them, and both rest on the same human-centred commitment that AI should support human judgement, never replace it. AILit, in effect, zooms DigComp in to our learners' age, and it is candid that it stops short of telling us how to assess what it describes.

The third question, the one AILit leaves open, is about **the task itself**, and the **AI Assessment Scale (AIAS)** (Perkins et al., 2024; adapted for EFL by Roe et al., 2025) answers it. The AIAS is a practical tool for deciding, and signalling on the task itself, how much AI a given assessment allows, across five non-hierarchical levels:

1. No AI: completed without AI, in controlled conditions.
2. AI Planning: AI for pre-task work only (brainstorming, research); the assessed work is independent.
3. AI Collaboration: AI on specific, defined sub-tasks; students critically evaluate and shape its output and keep their own voice.
4. Full AI: AI may complete elements of the task; the assessment is how well students direct and critique it.
5. AI Exploration: AI used creatively to address new problems, often co-designed with students.

No level is "better" than another, and the co-constructed traffic-light agreement from Section 1 is simply a class-friendly version of the same idea. Read together, the three frameworks form a single line of reasoning rather than a pile of acronyms. DigComp 3.0 names the competences a learner should build; AILit makes those competences concrete and age-appropriate; and the AIAS turns them into a design decision a teacher can make, and state, on every task. The first two tell us what is worth assessing. The third tells us how much of the tool belongs in the room while we do it.

5. Putting it to work: ten AI-aware English assessments

What follows is the practical core: two worked examples for each level of the AI Assessment Scale, designed for lower and upper secondary English. Each gives the task, the learners, why it stays valid, what you actually assess, and the competence it develops. They are starting points to adapt, not a complete catalogue.

Level 1, No AI: protect the baseline

You cannot interpret AI-assisted work unless you also have evidence of what a learner can do unaided. Level 1 is not nostalgia. It is the control condition that makes every higher level readable.

1.1 Human or machine? (lower to upper secondary, A2 to B2) Give students two short English texts on a familiar topic, one written by a person (drawn from a real source) and one generated by AI. In groups, and with no devices, they argue which is which, defending their verdict with linguistic evidence: register, specificity, cliché, the kinds of small errors humans and machines each make. The task is pure reading and speaking, and it doubles as critical AI literacy.

You assess: reasoning, the language of comparison and evaluation, and the quality of the evidence cited.

Competence link: this is almost word for word a DigComp 3.0 outcome, "recognise that it can be difficult to distinguish between information and content generated by humans and AI systems" (LO1.2.04), and AILit's "evaluate whether AI outputs should be accepted, revised, or rejected."

1.2 The live local anchor (lower secondary, A2 to B1) Students write a short narrative in class, by hand, from a prompt revealed only that day and tied to shared experience: a place in your town that matters to you, a moment from this week. Because the prompt is specific, personal, and unpredictable, and the writing happens in front of you, it evidences genuinely independent production.

You assess: spontaneous use of target tenses, coherence, and range, under known conditions.

Competence link: the teacher models that some tasks are deliberately No AI, and explains why (DigComp 3.0, Managing AI, "decide whether to use AI based on the nature of the task").

Level 2, AI Planning: let the machine help them start, assess what they build

Ideation and research are where AI genuinely helps. The move is to let it in at the front door while keeping the assessed product independent.

2.1 Brainstorm, then defend (upper secondary, B2) For an argumentative task on a relevant local or national issue, for example whether phones should be restricted in Portuguese schools, students first use AI to brainstorm arguments and vocabulary at home, submitting the chat log and a short note on which ideas they kept, dropped, or distrusted. They then write or debate the argument in class, unaided.

You assess: the independent piece, plus the quality of their selection and critique of the AI's ideas.

Competence link: AILit, Creating with AI ("use AI to explore new perspectives while staying accountable for the final content"); DigComp 3.0, Evaluating information.

2.2 Questions for a real voice (lower to upper secondary, B1) Before interviewing a family member about how English entered their life (school, work, music, migration), students prompt AI to generate guiding questions and a topic vocabulary bank. They then conduct the real interview, gather the answers, and present their findings.

You assess: the researched content and the spoken delivery, not the AI-made scaffold.

Competence link: DigComp 3.0, Managing AI (delegating structured, preparatory tasks); this also connects to the informal English our learners already absorb outside class.

Level 3, AI Collaboration: the richest level for English

Most future writing will be collaborative with AI. The validity move is to assess the student's decisions about the output, not the output itself.

3.1 Draft, feedback, revise, with a rationale (upper secondary, B1 to B2) Students write a first draft under Level 1 conditions. They then ask AI for feedback on one defined dimension (cohesion, or a specific target structure), revise, and submit a small portfolio: original draft, the AI feedback, the final version, and a short rationale explaining which suggestions they accepted or rejected, and why.

You assess: the revision decisions and the metalinguistic reasoning, where the learning is visible, rather than the polish of the final text.

Competence link: this mirrors AILit's own secondary scenario almost exactly, "use an AI writing assistant to revise a personal narrative, then reflect on whether its suggestions supported authentic voice."

3.2 Rehearse with AI, perform with a human (upper secondary, speaking) Students rehearse a speaking scenario with an AI conversation partner at home, then perform the unrehearsed equivalent live with a peer, and reflect briefly on what the AI helped with and where it fell short. This is the natural place to discuss, openly, the habit of treating a tireless, agreeable machine as if it were a person.

You assess: the live human performance and the reflection, not the private rehearsal.

Competence link: DigComp 3.0, Interacting through and with digital technologies, which explicitly asks learners to describe AI's role accurately and avoid anthropomorphism.

Level 4, Full AI: assess the orchestration, not the output

When AI may do the producing, the assessable skill becomes judgement: directing, evaluating, and correcting the machine. For language teachers this is unusually rich, because that judgement is itself deeply linguistic.

4.1 Translation post-editing (upper secondary, B2 and above) Give students an AI translation of a text (English to Portuguese, or the reverse) and have them post-edit it to publishable quality, annotating each change: false friends, register slips, idiom, mistranslation, tone. The machine produced the draft; the assessed skill is the critical linguistic evaluation that repairs it.

You assess: the accuracy and justification of their edits, a genuine and employable competence.

Competence link: AILit, "evaluate whether AI outputs should be accepted, revised, or rejected"; DigComp 3.0, Content creation.

4.2 Generate, then fact-check (upper secondary, B2) Students fully prompt AI to produce English content about Portuguese culture, history, or a local landmark, then critically annotate and fact-check it against reliable sources: where is it vague, generic, inaccurate, or quietly stereotyped? They justify a set of edits that would make it genuinely good.

You assess: the quality of their prompting and of their critique, orchestration and evaluation rather than authorship of the prose.

Competence link: DigComp 3.0, Managing AI and Evaluating information; AILit, Self and Social Awareness (bias and representation).

Level 5, AI Exploration: co-designed, creative, new

At the top of the scale, AI becomes material for genuine inquiry. These tasks are more open and project-shaped, and they return the most in motivation and in critical literacy.

5.1 Build a local language helper (upper secondary, project) Students design and prompt-engineer an AI assistant for a real audience: younger learners at their school struggling with a tricky grammar point, or visitors to their town who need English help. They test it, document where it works and where it fails, and present it.

You assess: the design reasoning, the iterative testing, and the spoken presentation.

Competence link: AILit, Designing and Managing AI ("describe how AI can be designed to support a community problem"); this loops back to localised materials creation.

5.2 Does AI speak our English? (upper secondary, co-designed inquiry) Working from a question the class defines together, students use AI to generate English about Portuguese varieties, registers, or culture, then fact-check and critique it: what does it get right, flatten, or stereotype? The exploration treats AI's limits as the object of study and produces a piece of critical, evidence-based English.

You assess: the critical analysis and the language used to express it.

Competence link: AILit, Creating with AI and Self and Social Awareness; DigComp 3.0, Evaluating information.

6. Design principles for AI-resilient assessment

The examples above are particular, but the moves behind them are general. Six design principles travel to almost any task an English teacher might set, and they are worth stating in their own right.

Assess the process, not only the product. This is the single most important shift, and it follows directly from the validity argument. If we grade only a finished text, we grade something a machine can now produce; if we mark the drafts, the notes, the decisions, and the reflections that led to it, we grade the learning itself, which a machine cannot do on the student's behalf (Dawson et al., 2024). In practice this means building visible thinking into the task: a planning stage that is collected, a revision history that is annotated, a short account of the choices made along the way. The product still matters, but it becomes the end of a trail of evidence rather than the only thing we see. For writing in particular, this turns the very feature that makes AI threatening, its ability to hand back a polished paragraph, into a non-issue, because the paragraph is no longer where the marks are.

Anchor tasks in the personal and the local. When a prompt is rooted in a learner's own experience, their street, their family, a debate the class had yesterday, an AI-generated answer loses much of its power, because the relevant material lives in the student and in the room, not on the internet. Authenticity is not decoration here; it is a validity strategy.

Be safe and open. A useful way to organise a whole unit, rather than a single task, is to think in two

lanes. One lane is safe: in-class, supervised work, often at No AI or AI Planning, that confirms what a learner can do unaided and gives every higher-level task something to be read against. The other lane is open: AI-permitted, authentic tasks that look like the real world and deliberately build responsible use. Some universities now formalise exactly this split, pairing a secured task with an open one across a programme so that integrity and authenticity are handled by different tasks rather than fought over within one (University of Sydney, 2024). The AI Assessment Scale is the natural companion to this thinking, because it lets you place each task in its lane and say so plainly.

Insist on oral and face to face interactions. A short spoken defence of a written piece, an in-class interview, a real-time discussion, these assess constructs that a chatbot simply cannot perform for a student: interaction, mediation, listening, thinking on one's feet. They need not be high-stakes. Two minutes of follow-up questions on a submitted essay will tell you, quickly and fairly, how much of it the student can stand behind.

Make reflection part of the task. Asking learners to explain, briefly, how they approached a piece of work and what they decided about any AI they used is not an add-on; it is itself a valid and assessable object, and it is one of the simplest ways to make reasoning visible.

Diversify formats, and always signal the AI-use level. A podcast, a poster, an annotated draft, a recorded presentation, a sequence of small in-class tasks: varied modes are harder for AI to mimic wholesale and they reach competences a single essay never touches. Whatever the format, state the AIAS level, or the traffic-light colour, on the task itself, so that the rule carries a purpose the class already understands and a teacher can hold students to.

7. What this approach is not

Because balance matters more than slogans, five clarifications are worth making explicit.

It is **not surveillance**. None of these tasks depend on catching anyone, and none require a detector. They depend on designing work in which genuine understanding is what gets measured, which makes the question of who used what beside the point. A classroom built on detection breeds suspicion; a classroom built on valid tasks does not need it.

It is **not anti-AI**. Only two of the ten examples exclude AI, and they do so to protect the baseline that makes the other eight readable. The whole logic of the scale is that No AI and Full AI are equally legitimate choices for different goals, and the higher levels deliberately teach students to use these tools well, because they will have to.

It is **not one size fits all.** Level and proficiency matter. A B2 class can post-edit a translation; an A2 class needs the live anchor first. The scale is a planning lever, not a ladder to climb, and a teacher will move up and down it within a single unit depending on what each task is meant to reveal.

It is **not a one-off fix**. Redesigning assessment for AI is not a box to tick once and forget. The tools change, learners' habits change, and what counts as a secure or an authentic task will keep shifting. The value of working from frameworks and a clear principle, rather than from a list of banned behaviours, is precisely that they hold steady while the particulars move.

It is **not a lowering of standards**. AI-aware assessment does not ask less of students. It asks something harder, and more worth knowing.

8. Closing

The point of an AI-aware assessment is not to discover whether a student used a machine. It is to design work in which we can still see the person think. That is the question worth assessing for, and it is the one our students, the first to grow up surrounded by AI, most need us to keep asking. The question is no longer whether they used AI, but whether we can still see them think.

References

Bennett, S., Maton, K., & Kervin, L. (2008). The "digital natives" debate: A critical review of the evidence. *British Journal of Educational Technology, 39*(5), 775-786. https://doi.org/10.1111/j.1467-8535.2007.00793.x

Cosgrove, J., & Cachia, R. (2025). *DigComp 3.0: European Digital Competence Framework (Fifth edition)*. Publications Office of the European Union. https://doi.org/10.2760/0001149

Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education, 49*(7), 1005-1016. https://doi.org/10.1080/02602938.2024.2386662

Lindade, C. (2026). From bans to balance: Co-creating AI policies with students in the EFL classroom. *APPI eJournal, (28)*, 7-10. https://sigarra.up.pt/flup/pt/pub_geral.pub_view?pi_pub_base_id=778393

McCrindle, M., & Fell, A. (2021). *Generation Alpha: Understanding our children and helping them thrive*. Hachette Australia.

OECD. (2025). *Empowering learners for the age of AI: An AI literacy framework for primary and secondary education* (Review draft). OECD Publishing. https://ailiteracyframework.org

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for the ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice, 21*(6), Article 6. https://doi.org/10.53761/q3azde36

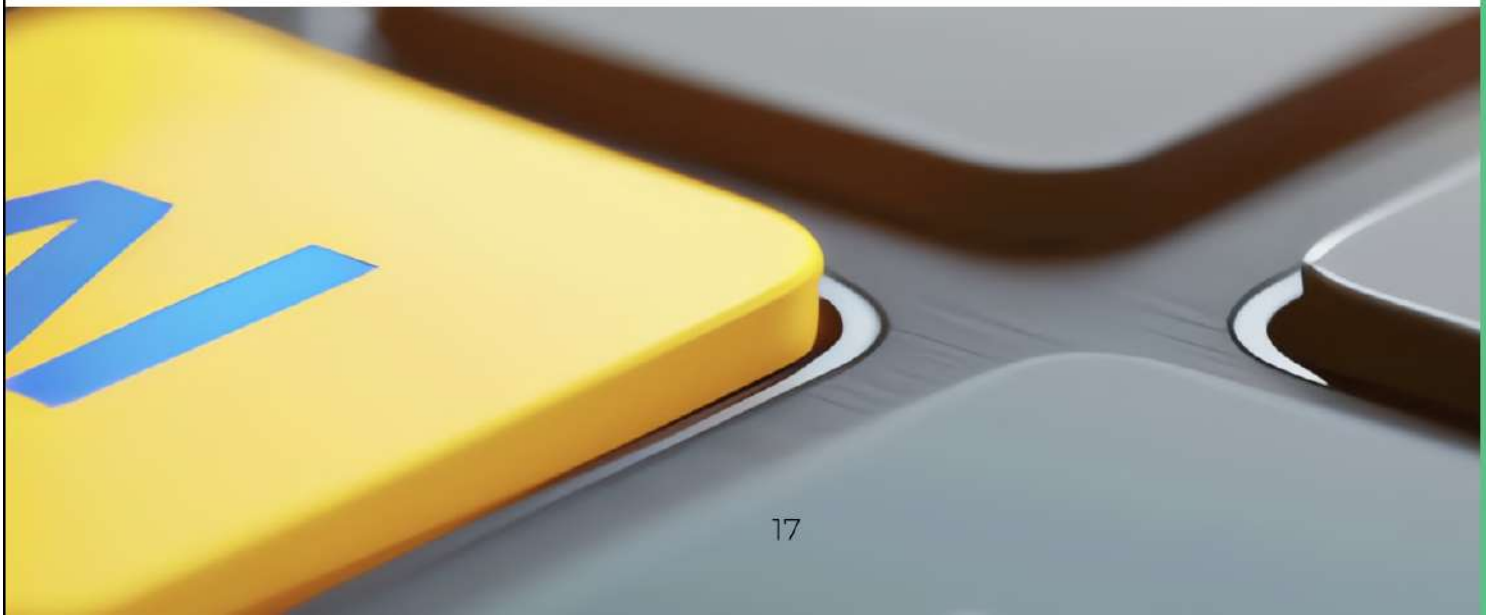
Roe, J., Perkins, M., & Furze, L. (2025). *From assessment to practice: Implementing the AIAS framework in EFL teaching and learning* [Preprint]. arXiv. https://arxiv.org/abs/2501.00964

Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis*. Appleton-Century-Crofts.

University of Sydney. (2024). *The two-lane approach to assessment in the age of AI*. Teaching@Sydney. https://educational-innovation.sydney.edu.au/teaching@sydney/frequently-asked-questions-about-the-two-lane-approach-to-assessment-in-the-age-of-ai/

Vodafone Foundation. (2024). *AI in European schools: A European report comparing seven countries*. Vodafone Foundation. https://skillsuploadjr.eu/docs/contents/AI_in_European_schools.pdf

Carlos Lindade, APPI member B-7249, is a Portuguese Canadian ELT professional heavily involved in training future EFL teachers. He holds a PhD in Advanced English Studies from the University of Vigo and is an Assistant Professor at the Faculty of Arts and Humanities of the University of Porto (FLUP). He is an integrated researcher of the Centre for English, Translation and Anglo-Portuguese Studies (CETAPS) and a regular speaker at APPI events.



appi

ASSOCIAÇÃO PORTUGUESA
DE PROFESSORES DE INGLÊS

ISSN 2184-7525