

FORMATIVE ASSESSMENT AND AI IN ENGLISH-LANGUAGE TEACHING

Younger teachers report doing more formative assessment. The AI fix for the gap is still unproven.

New survey evidence from English-language classrooms in Almaty, Kazakhstan, translated for teachers elsewhere weighing up AI-assisted formative assessment.

Before you add an AI tool to your marking routine, how would you know if it actually saved you time?

In one line

A survey of 60 Kazakhstani English teachers found younger teachers reporting far higher use of every formative-assessment practice, including digital and AI tools, than colleagues aged 46 and over. The researchers argue AI could help close that gap, but they never tested whether it actually does.



Every formative-assessment habit was more common among younger teachers, self-reported.

60 EFL teachers in public schools across Almaty city and region, Kazakhstan, split into three age bands (24-34, 35-45, 46+, 20 teachers each) completed a survey on their formative-assessment practice; 12 were then interviewed in depth. On every practice the survey asked about, from sharing lesson goals to running peer-assessment, teachers aged 24-34 reported using it more than teachers aged 46 and over. Time constraints, large classes and lack of training, not lack of interest, were the barriers teachers named most, across every age band.

60

EFL teachers, Almaty: 3 age bands of 20, plus 12 follow-up interviews

40% vs 15%

reported any use of digital or AI tools, youngest vs oldest teachers

✓

Immediate in-class checks (e.g. quick quizzes) · 90% (24-34) vs 50% (46+)

✓

Peer-assessment activities, occasionally or more · 55% (24-34) vs 20% (46+)

✓

Regular learner self-assessment · 50% (24-34) vs 30% (46+)

•

No chi-square or ANOVA result is reported for any of these gaps · Both tests are named as the method, but no statistic appears in the results

Who reported using a digital or AI tool for formative assessment?

Share of teachers in each age band who reported any use, including simple things like online quizzes or automated feedback.



Self-reported survey data, not classroom observation. No formal test of whether this gap is statistically reliable is reported in the paper.

What “AI-assisted formative assessment” actually meant in this paper

FA COMPONENT	WHAT IT SOUNDS LIKE	WHAT THE PAPER ACTUALLY SHOWS
Immediate instructional adjustment	<i>“Quick poll after teaching possessive apostrophes: most learners pick the wrong answer, so the teacher pauses and re-explains before moving on.”</i>	One teacher’s own classroom story in this paper describes exactly this, using a free polling app. It is a single example the authors chose to illustrate their argument, not evidence collected across the 60-teacher sample.
SL GUIDANCE Any quick-poll tool works for this. The mechanism is the pause-and-reteach step, not the software.		
Feedback	<i>“A grammar checker flags an error and suggests a fix before a learner submits the final draft.”</i>	Teachers who used Turnitin’s Feedback Studio told the researchers it saved time and made comments more consistent. That is teacher-reported experience, not a measured before-and-after result.
SL GUIDANCE Time saved on checking mechanics is the credible claim here, not any effect on writing quality, which this paper did not test.		
Self-assessment	<i>“A learner checks their own paragraph against a shared rubric before handing it in.”</i>	The paper’s account of AI-assisted self-assessment is drawn entirely from other researchers’ work and general description, not from this study’s own 60 teachers.
SL GUIDANCE Introduce the rubric on paper first. Only add a tool once learners can already judge their own work against it.		
Peer-assessment	<i>“My students don’t want to criticize their friends, so they either give all high marks or avoid saying anything useful.”</i>	A teacher’s own words in this study. Peer-assessment was the least-practised FA component overall, held back by learners’ reluctance to critique classmates, not by lack of a digital tool.
SL GUIDANCE Anonymous digital peer review, via any classroom platform, may ease this reluctance more reliably than adding an AI element on top.		

The honest fence: where this evidence stops

No significance test reported

Chi-square tests and ANOVA are named as the method for comparing age groups, but no statistic or p-value for any age-group gap appears anywhere in the results. Treat the gap as a self-reported difference, not a confirmed effect.

The AI claim is untested by design

The authors state directly, in their own limitations section, that AI's effect on learning was not experimentally tested and the suggested approaches "still need further validation."

Almaty only

The sample is drawn from one city and its surrounding region. The authors say this may not represent the whole country, especially rural or non-Russian-speaking areas of Kazakhstan.

Self-reported, not observed

Teachers reported what they do, not what a classroom observer recorded them doing. The authors themselves flag that self-reported data may include bias.

The clearest AI examples are single-teacher stories

The quick-poll story and the pronunciation-app story are each one classroom's experience, offered as illustration, not data collected systematically across the 60-teacher sample.

One borrowed statistic isn't about these teachers

An 89% figure on self-assessment quoted in the paper comes from a different, external study (Mazloomi & Khabiri, 2018), not from the 60 Kazakhstani teachers this finding is about.

● Early signal

A real survey with a genuine practice gap by teacher age, but no significance testing reported, and a central AI-effectiveness case built on citations to other research and single-teacher anecdotes rather than a measured test, by the authors' own admission. Reported test statistics for the age comparisons, a wider sample beyond Almaty, and an actual before-and-after trial of one AI tool would raise this grade.

Worth testing: does an AI tool actually save you time, or does it just feel new?

One reading: this looks less like an appetite gap between generations and more like a time-and-training gap. Every age band named the same barriers, large classes, limited time, and limited training, regardless of how comfortable they were with technology. If that is right, an AI tool is only worth adopting because it demonstrably saves time or catches something you would have missed, not because it is new.

1 Trial one tool, one task

Pick a single formative-assessment task you already do by hand, such as a quick check for understanding, and trial one free digital or AI tool for it, rather than adopting several tools at once.

2 Measure time, not novelty

Track marking time saved and misconceptions caught over two weeks. Neither was measured for any AI tool in this paper, so your own two weeks of notes will tell you more than the study can.

3 Keep the process step, not just the tool

The paper's clearest example worked because the teacher paired the tool with an immediate follow-up explanation. The software surfaced the gap; the teacher closed it.

Early signal, so run it as an experiment: for two weeks, use one quick digital or AI check, such as an online quiz or poll, for a single formative-assessment task you already do by hand. Note whether it saves you marking time or catches a misconception you would have missed, not whether it feels new.

Why not try? Planning a two-week formative-assessment trial is easier with a structure to build it from. Open a free account to plan your next check for understanding at app.structural-learning.com.