**Journal of English Language Teaching and Applied Linguistics**
ISSN: 2707-756X
DOI: 10.32996/jeltal
Journal Homepage: www.al-kindipublisher.com/index.php/jeltal

JELTAL

AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

**| RESEARCH ARTICLE**

# Genetic Algorithm Versus Random Feedback Selection in Computer-Assisted Language Learning

**Yassine Khaya**

English Department, Ibn Tofail University, Kenitra, Morocco

**Corresponding Author**: Yassine Khaya, **E-mail**: khaya.yassine@gmail.com

**| ABSTRACT**

This study tests whether the rule used by a computer-assisted language learning (CALL) tool to select corrective-feedback types affects learners' English-tense accuracy. A browser-based tutor delivered one of six feedback types after each error. In one condition, a genetic algorithm assigned feedback using learner-specific value estimates, preference weights, and a repetition penalty; in the other, types were assigned randomly, while items, order, message count, and total feedback time were matched. Of 255 Moroccan university students allocated, 103 completed five sessions. The algorithm changed the feedback mix, but accuracy did not differ between conditions (odds ratio = 0.90, 95% confidence interval [0.67, 1.20]; 60.2% vs. 57.4%) and did not increase across sessions. Logs indicated time pressure for many messages, weak preference weights, and value estimates partly driven by item distribution. Adaptive feedback selection requires valid learner signals and adequate processing time, not merely a different assignment rule.

**| KEYWORDS**

corrective feedback; adaptive CALL; feedback selection; genetic algorithm; English tense; controlled comparison

**| ARTICLE INFORMATION**

**ACCEPTED:** 30 August 2026 **PUBLISHED:** 27 September 2026 **DOI:** 10.32996/jeltal.2026.8.10.1

## 1. Introduction

Corrective feedback is among the most widely provided and least systematically designed features of CALL tools. When a learner chooses the wrong verb form, a tool can show the correct form, restate the sentence, state the rule, point to the cue in the sentence, ask a question about the form, or attach a label naming the error. Second language acquisition (SLA) research has studied these options mainly as choices a teacher makes in real time (Lyster & Ranta, 1997; Lyster et al., 2013). In a tool the responses are written in advance, several are available for the same error, and something in the code has to decide which one to display. In most tools that decision is a fixed rule written by the developer, and whether a rule that adapts to the learner would serve learners better represents a significant yet underexplored question in CALL research.

The present study addresses this question through a controlled comparison of two decision rules embedded in the same practice tool. A learner who completed a round of 32 tense items and made, for example, eleven errors was then shown eleven short messages, one per error, for a total of 90 seconds. Each message was of one of six types, and the time on screen was shared among the messages according to a weight the system held for each type and each learner, so that the type given to an error also decided how long its feedback stayed visible. In one condition a genetic algorithm, explained in Section 2.3, chose the type for each error; in the other the type was chosen at random. The intended experimental difference was the assignment rule; an unintended difference in message ordering is reported in Section 5.5.

The scope is deliberately narrow. The study does not ask whether corrective feedback helps learners of English tense, a question on which the literature offers a qualified affirmative answer (Li, 2010; Russell & Spada, 2006; Van der Kleij et al., 2015). It asks what a controlled evaluation can establish about an adaptive assignment rule once a tool already offers several types. Two considerations make this worth a test. Adaptive selection of instructional material is among the main promises of artificial intelligence in language education, yet reviews consistently find that adaptive rules are seldom tested against active baselines with equivalent exposure, and seldom accompanied by any check that the rule changed what learners experienced (Doroudi et al., 2019; Huang et al., 2022; Liang et al., 2023; Tan et al., 2025). And the tutorial CALL literature indicates that the type of

feedback a system gives affects whether learners take it up (Heift, 2004, 2010; Nagata, 1993), which makes the assignment of types a pedagogical variable rather than a technical detail.

The findings are largely negative, and the contribution lies elsewhere: in an evaluation design that separates whether the algorithm changed the intervention from whether the changed intervention improved performance, and in a diagnosis, drawn from the logs and the source code, of why a working algorithm produced no advantage. Section 2 sets out the background and the research questions, Section 3 the system and the analysis, Section 4 the results, and Section 5 the interpretation and its implications.

## 2. Literature Review

### 2.1 Corrective Feedback in Tutorial CALL

Corrective feedback (CF) refers to responses that signal to a learner that their language contains an error (Nassaji & Kartchava, 2021). Classroom research has classified CF into recasts, explicit correction, metalinguistic feedback, elicitation and repetition, and linked some of these to uptake and later development (Lyster & Ranta, 1997; Lyster & Saito, 2010). Meta-analyses agree that CF has a moderate positive effect overall, while indicating that its size depends on the target structure, the learners' proficiency, the type of feedback and the outcome measure (Kang & Han, 2015; Li, 2010; Russell & Spada, 2006), and Truscott (1996) remains a reminder that the effect is not guaranteed. In tutorial CALL the feedback is generated by the system and delivered as text on a screen (Heift & Schulze, 2007, 2015), and the evidence concerns mainly how explicit it should be: feedback that explains the rule outperforms feedback that only flags the error (Nagata, 1993; Van der Kleij et al., 2015), and metalinguistic explanations produce more repair attempts than error highlighting, especially among less proficient learners (Heift, 2004, 2010). Reviews of formative feedback add that feedback must be specific, timely and actually processed in order to have any effect (Hattie & Timperley, 2007; Shute, 2008). Two points follow for the present study. The type of message influences what learners do with it, and more explicit messages take longer to read, and learners do not always read them (Ranalli, 2018). A tool that displays feedback for a limited time therefore has to decide not only which type to show but how much time each message receives, and the second decision can undo the first.

### 2.2 The Selection Problem in Adaptive CALL

When a system can produce several kinds of feedback for the same error, it must decide which to display and for how long. Adaptive systems make that choice depend on the learner or on evidence about what has worked; in intelligent tutoring research this is instructional sequencing, approached with procedures from mastery rules to reinforcement learning (Clement et al., 2015; Doroudi et al., 2019; VanLehn, 2011). The record is sobering. Doroudi et al. (2019) found that data-driven sequencing outperformed a baseline in only about half of the comparisons reviewed, and often only when the baseline was weak. Two flaws recur: comparison with a no-feedback condition, which confounds the rule with exposure, and the absence of any check that the rule changed the learning experience, so that a null result cannot be distinguished from a rule that did nothing. Reviews of artificial intelligence in language education report the same pattern (Huang et al., 2022; Kartal & Yeşilyurt, 2024; Liang et al., 2023; Tan et al., 2025). The present study uses an active control with matched exposure, and establishes whether the algorithm changed what learners received before asking whether it changed how they performed.

### 2.3 The Approach Adopted Here

After each round the tool faced a fixed problem: each error would receive one message; the messages would share a time allowance set by the number of errors; and the allowance would be divided among them in proportion to a learner-specific weight for each type. Choosing a type for an error therefore also decided how much of the fixed time that error's feedback would receive, at the expense of the others. The problem resembles a packing problem with a fixed capacity and far more possible assignments than can be checked. A genetic algorithm (GA) is a general-purpose search method for such problems (Goldberg, 1989; Holland, 1975; Mitchell, 1998): it starts from random candidate assignments, scores them, keeps and recombines the better ones with small random changes, and repeats for a fixed number of rounds. GAs have been used in education mainly for assembling tests under several constraints at once (Verschoor, 2007).

Each candidate assignment was scored on three criteria, specified in Section 3. *Value*: a learner-specific estimate, for each feedback type and tense category, of whether errors in that category had decreased after that type was shown. *Preference weight*: a learner-specific weight for each type, built from how long the learner had read messages of that type during a placement test and from survey ratings, which also set each message's share of the time allowance; the use of reading time as an indicator of engagement follows educational measurement (Wise & DeMars, 2006; Wise & Kong, 2005), and the assumption that formats differ in processing demand follows Cognitive Load Theory (Paas & van Merriënboer, 1993; Sweller, 1988). *Variety*: each further use of a type in a round counted for half as much as the previous, on the finding that varied practice supports retention (Bjork & Bjork, 2011; Schmidt & Bjork, 1992). The design reflects Feedback Intervention Theory, in that all six types address the

specific error and the value estimates were meant to favour whichever type the learner responded to (Kluger & DeNisi, 1996). This is a rationale, not a prediction that the GA would win; the literature above gives ample reason to expect small differences between types under short, text-only conditions. English tense and aspect was chosen because it is frequent, remains difficult into intermediate proficiency (Bardovi-Harlig, 2000), suits an item format in which each type can be generated from a template, and maps unevenly onto the learners' other languages, Moroccan Arabic, Standard Arabic and French (Boubekri & Ech-Charfi, 2023; El Hadari, 2022).

### 2.4 Research Questions

RQ1. What does a controlled comparison with an exposure-matched active baseline reveal about the contribution of an algorithmic rule for assigning feedback types to learners' accuracy on English tense items, over and above what the practice format itself produces?
RQ2. What do the delivery data logged by the system, that is the composition of the feedback sets, the time each message was displayed, and the learner-specific inputs on which the rule depended, reveal about the conditions under which such a rule can make a difference, and what follows for the design and evaluation of adaptive corrective feedback in CALL?

Two expectations stated in the protocol are retained as hypotheses. H1: feedback sets assigned by the GA would be more varied than those assigned at random, the condition under which RQ1 is answerable at all. H2: accuracy would rise across sessions in both conditions, more steeply in the GA condition if the value and preference estimates captured useful structure.

## 3. Methodology

The system has several parts, each depending on the one before it. This section follows the order in which a participant met them and introduces each part before it is used.

### 3.1 The Participant's Path

A participant registered and gave consent (stage 1 in Figure 1), took a *placement test* in which every wrong answer was followed by a feedback message they dismissed with a button (stage 2), rated the six kinds of message in a *survey* (stage 3), and from their reading times and ratings received a *preference weight* for each kind (stage 4). They were then allocated to a condition (stage 5) and completed five *practice sessions* of two *rounds* of 32 items (stage 6). After every round came a *feedback phase* (stage 7): every error received one message, the messages shared a fixed *time allowance* set by the number of errors, and the allowance was divided among them by the weights. The intended experimental difference was the rule that chose the kind of message each error received. The algorithm drew on *value estimates* updated once per session (stage 8). Table 1 fixes the terms.

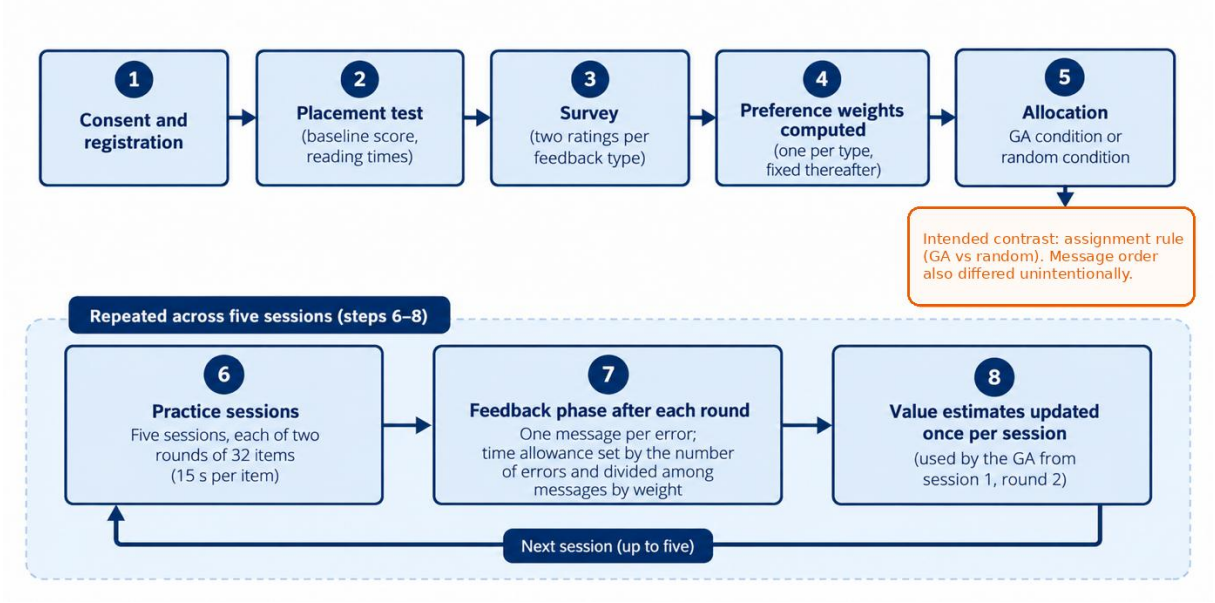


**Figure 1**
*The participant's path through the study. Stages 1 to 4 happen once; stages 6 to 8 repeat over five sessions. The two conditions differ at the point marked in orange (and, unintentionally, in the order in which messages were shown; Section 3.6).*

**Table 1**
*Terms used in this article*

| Term | Meaning in this study |
|---|---|
| Item; round; session | A sentence with a gap and four verb forms to choose from; 32 items followed by a feedback phase; two rounds. Five sessions in all. |
| Error | An item answered with a wrong option. Unanswered items are not errors. |
| Feedback type; message | One of six kinds of message (reformulation, direct correction, brief explanation, elicitation, location, error code); the text shown for one error, of one type. One message per error. |
| Time allowance | The seconds the messages of a round share on screen; fixed by the number of errors (Section 3.6). |
| Preference weight (W) | A number between 0 and 1 for each type and each participant, from placement reading times and survey ratings; the six sum to 1 (Section 3.4). |
| Value estimate (V) | A number for each participant, type and tense cell, starting at 0.5, recording whether errors in that cell fell or rose after that type was shown (Section 3.8). |
| Tense cell | One of twelve combinations of tense category (simple, continuous, perfect, perfect continuous) and time reference (present, past, future). |

### 3.2 Design, Allocation, Participants

The two conditions shared the same item bank, item order, six feedback types, one-message-per-error rule and time-allowance rule; the random condition is therefore an active control with matched exposure; the intended experimental contrast was the assignment rule. Allocation took place when the researcher approved a participant who had completed the placement test and the survey: the application assigned each newly approved participant to whichever condition had fewer members, choosing at random when the two were equal. This balanced alternation is why the study is reported as a controlled comparison rather than a randomised trial; participants could not see or influence it, but the sequence was predictable from the order of approval (Section 5.5). Reporting follows CONSORT as far as the data allow (Schulz et al., 2010).

The study received fast-track ethical approval from the Department of Computer Science Ethics Committee, University of York (approval reference: Khaya20250729). Participants were undergraduate students of English at a public university in Morocco; participation was voluntary and unpaid, each participant ticked consent statements including that they were over eighteen, and registration required only a name and an email address, removed before analysis, so that age, gender and year of study cannot be reported. Two researcher test accounts were removed from all counts. In total 683 people consented between 13 October and 10 December 2025, 328 completed the placement test, 255 completed the survey and were allocated (127 GA, 128 random), and 103 completed all five sessions.

### 3.3 Items and Feedback Types

The bank contained 320 multiple-choice items, 64 per session and 16 per session for each of the four tense categories, with an uneven distribution of time reference (180 present, 76 past, 64 future); every stem carried a time cue that determined the required form. The author wrote the stems, four options and, for each of the three wrong options, one message of each of the six types. All participants answered all 320 items once, in the same order. Table 2 shows the six types and the messages written for one item; the elicitation type went on to explain why each distractor was wrong and, because the display advanced automatically, could not be answered; the error code was shown as the internal label, without a key. Across the 12,536 messages delivered to completers, median displayed length was 14 words for brief explanations, 22 for reformulations, 23 for location cues, 28 for direct corrections, 59 for elicitation messages and one token for codes.

**Table 2**
*The six feedback types, their counterparts in the CF literature, and the messages for the item "Sam ___ class yesterday" (options: has not attended, did not attend, was not attending, does not attend)*

| Type | Counterpart | Message as displayed | Words |
|---|---|---|---|
| Reformulation | Recast | Sam did not attend class yesterday. This reformulation shows how the time expression "yesterday" determines the tense/aspect. | 17 |

| Type | Counterpart | Message as displayed | Words |
|---|---|---|---|
| Direct correction | Explicit correction | The correct answer is "did not attend" because form past simple negative with "did not" + base verb when a clear past time is given. The sentence context triggers this form. | 31 |
| Brief explanation | Metalinguistic feedback | Use Past Simple with definite past-time adverbials like "yesterday". | 9 |
| Elicitation | Elicitation with comment on distractors | What does the phrase "yesterday" imply about when the action happens? Which form encodes that meaning? Also, consider why the distractors do not match: "does not attend" Present Simple mismatches the definite past-time cue; "was not attending" Past Continuous depicts an ongoing past situation, not a simple non-occurrence; "has not attended" Present Perfect generally avoids definite past-time adverbs like "yesterday". | 60 |
| Location | Indirect feedback | Focus on the cue "yesterday" in the stem. It tells you to adjust the blanked verb to match the required tense/aspect here. | 22 |
| Error code | Coded feedback | PAST.SIMP.NEG | 1 |

### 3.3.1 Two Items From Each Tense Cell

Items are shown as stored; on screen the gap appeared as a blank, the bracketed base form was shown, and the four options were lettered A to D in the order stored. The correct option is italicised here.

**Table 3**
*Example items*

| Tense cell | Item | Stem | Options |
|---|---|---|---|
| Simple, present | S1R1-SIM-001 | Maria ___ (visit) her grandmother every Sunday afternoon. | visited; is visiting; *visits*; has visited |
| Simple, present | S1R1-SIM-002 | He always ___ (eat) breakfast before school. | will eat; ate; is eating; *eats* |
| Simple, past | S1R1-SIM-003 | Yesterday, the meeting ___ (end) early. | was ending; has ended; ends; *ended* |
| Simple, past | S1R1-SIM-005 | They ___ (not go) to the museum last weekend. | have not gone; *did not go*; were not going; do not go |
| Simple, future | S1R1-SIM-004 | I ___ (call) you later tonight, I promise. | call; am calling; called; *will call* |
| Simple, future | S1R1-SIM-007 | I think it ___ (rain) tomorrow. | rains; has rained; is raining; *will rain* |
| Continuous, present | S1R1-CON-009 | She ___ (cook) dinner right now. | *is cooking*; was cooking; cooks; will be cooking |
| Continuous, present | S1R1-CON-010 | He ___ (not work) this week because his car is in the shop. | will not work; *is not working*; does not work; was not working |
| Continuous, past | S1R1-CON-012 | They ___ (study) when the fire alarm went off. | *were studying*; studied; had studied; are studying |
| Continuous, past | S1R1-CON-013 | At 8 p.m. last night, I ___ (clean) the kitchen. | am cleaning; cleaned; *was cleaning*; will be cleaning |
| Continuous, future | S1R1-CON-014 | This time tomorrow, we ___ (fly) to Tokyo. | are flying; *will be flying*; will fly; flew |
| Continuous, future | S1R1-CON-015 | He ___ (wait) for you when you arrive. | waited; is waiting; was waiting; *will be waiting* |
| Perfect, present | S1R1-PER-017 | She ___ (finish) her homework already. | finished; *has finished*; was finishing; has been finishing |
| Perfect, present | S1R1-PER-018 | We ___ (live) in this city for ten years. | *have lived*; are living; have been living; lived |

| Tense cell | Item | Stem | Options |
|---|---|---|---|
| Perfect, past | S1R1-PER-020 | They ___ (leave) by the time I got there. | have left; left; were leaving; *had left* |
| Perfect, past | S1R1-PER-021 | She ___ (not see) the email before the meeting started. | was not seeing; has not seen; did not see; *had not seen* |
| Perfect, future | S1R1-PER-022 | By next Friday, I ___ (complete) the draft. | *will have completed*; will complete; have completed; will be completing |
| Perfect, future | S1R1-PER-023 | They ___ (finish) the work by the time you arrive tomorrow. | have finished; will finish; had finished; *will have finished* |
| Perfect continuous, present | S1R1-PCO-025 | I ___ (study) for three hours now. | *have been studying*; studied; am studying; had been studying |
| Perfect continuous, present | S1R1-PCO-026 | He ___ (train) since last summer. | is training; trained; will have been training; *has been training* |
| Perfect continuous, past | S1R1-PCO-027 | She ___ (run) all morning before she finally rested. | was running; has been running; *had been running*; ran |
| Perfect continuous, past | S1R1-PCO-028 | They ___ (work) for five hours when they decided to stop. | have been working; *had been working*; worked; were working |
| Perfect continuous, future | S1R1-PCO-029 | By next month, I ___ (work) here for a year. | will be working; *will have been working*; worked; have been working |
| Perfect continuous, future | S1R1-PCO-030 | By the end of this year, he ___ (study) English for three years. | will be studying; *will have been studying*; has been studying; studied |

### 3.4 Placement Test, Survey and Preference Weights

The placement test presented up to 60 items drawn at random from the bank; after each error one message was displayed, of a type chosen at random among those not yet shown three times, and the time until the participant dismissed it was recorded. The test ended at the first error after all six types had been shown three times (92% of participants at exactly 19 errors) or after 60 items. Its proportion correct is the baseline measure (M = 0.41, SD = 0.17 among allocated participants); its items came from the practice bank, so each participant had already met about a tenth of the practice items (Section 5.5). The survey then presented each type with a description and an example and asked for two ratings on a five-point scale: "This feedback style requires significant mental effort for me to process and understand" and "It takes my tutor a long time to deliver this feedback to me"; the system used the second as a rating of effectiveness, although as worded it concerns delivery time.

#### 3.4.1 The Feedback Survey

Each of the six feedback types was presented on its own page with the name and description below, an example built on the item "They ___ (visit) the museum yesterday" with the wrong choice "are visiting", a short teacher-student dialogue illustrating the type, and two statements, each rated on a five-point scale labelled Strongly Disagree to Strongly Agree: (1) "This feedback style requires significant mental effort for me to process and understand." (2) "It takes my tutor a long time to deliver this feedback to me."

**Table 4**
*Descriptions shown to participants*

| Type (name shown) | Description shown |
|---|---|
| Direct Correction | This feedback style gives you the correct answer immediately and explains why it's right. The teacher tells you exactly what the answer should be and provides a clear grammatical explanation. This is like having a tutor who says 'Here's the right answer and here's why it's correct.' |

| Type (name shown) | Description shown |
|---|---|
| Location-Based Hints | This feedback doesn't give you the answer directly. Instead, it points you to specific clues in the question that should help you figure it out yourself. The teacher guides your attention to important words or phrases that contain the key information. |
| Grammatical Codes | This feedback uses short technical codes to tell you what type of grammar mistake you made. It's like a quick label that identifies the specific grammatical category. You need to know what these codes mean to understand the feedback. |
| Brief Explanations | This feedback gives you a short, clear rule about when to use a particular grammar form. It's like a quick reminder of the grammar rule without giving you the specific answer. You need to apply this rule to figure out the correct form yourself. |
| Sentence Reformulation | This feedback shows you the complete correct sentence and explains how the context (the situation or time words) determines which grammar form to use. It's like seeing the full picture of how grammar works in context. |
| Elicitation | This feedback asks you questions to help you think through the problem yourself. Instead of telling you the answer, the teacher guides you with questions that lead you to discover the correct form and understand why it's right. It's like having a conversation that helps you learn. |

Figure 2 shows how the two sources became one weight per type. Mean reading time was converted to a score of 1 + 4t/15 capped at 5, and the whole-number part of the score, rescaled to 0 to 1, was the *behavioural weight*; any mean under 3.75 seconds therefore gave zero. The *survey weight* was the mean of the rescaled second rating and one minus the rescaled effort rating. The final weight was 0.6 times the behavioural plus 0.4 times the survey weight, and the six were normalised to sum to one and fixed for the participant. A full worked example is given in Sections 3.4.2–3.8.1.

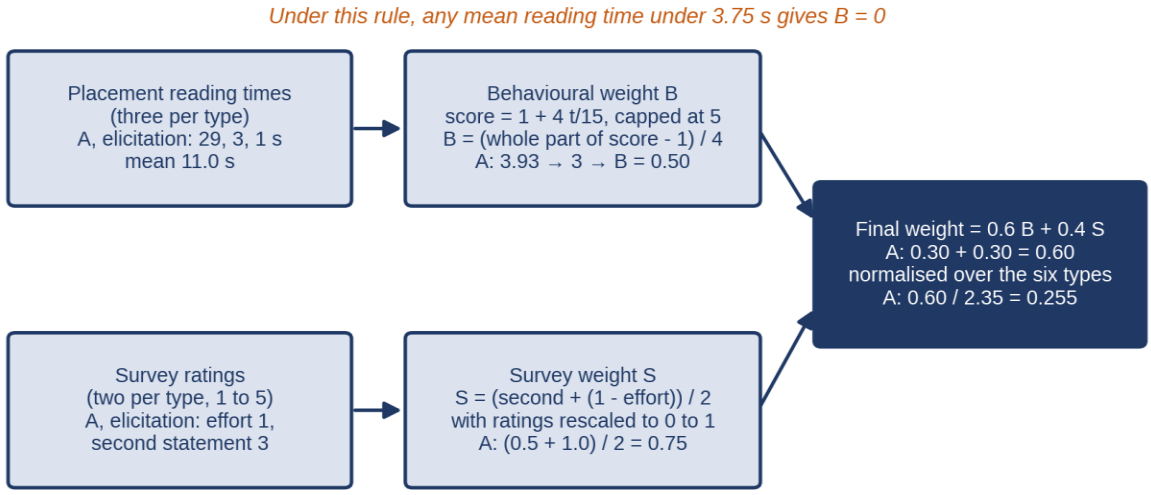


**Figure 2**
*How a preference weight was built, with Participant A's elicitation weight as the example.*

### 3.4.2 Participant A's Placement Reading Times and Survey Ratings

Participant A is one real participant in the GA condition; the pronoun "they" refers to that one person. The examples below use their logged numbers.

Participant A answered 25 of 44 placement items correctly, a proportion of 0.57. Their three reading times for each type, in seconds, were: elicitation 29, 3 and 1 (mean 11.0); error code 23, 3 and 1 (mean 9.0); direct correction 5, 9 and 2 (mean 5.3); brief explanation 7, 7 and 1 (mean 5.0); location 6, 1 and 2 (mean 3.0); reformulation 1, 1 and 2 (mean 1.3). The pattern is typical: a long first look at a new type, then one or two seconds for the next two, so that the means depend heavily on a single exposure. On the survey they rated the effort statement at 1 for elicitation, brief explanation and reformulation, at 2 for error code and direct correction, and at 3 for location; they rated the second statement at 3 for elicitation, error code, direct correction and location, at 2 for reformulation, and at 1 for brief explanation.

### 3.4.3 Preference-Weight Calculation

The two sources were combined into one preference weight per type in three steps, shown for Participant A in Table 5. First, the mean reading time t (in seconds) was converted to a *behavioural weight* B: the application computed a preference score s of 1 + 4t/15, capped at 5, and the server rescaled the whole-number part of s to the range 0 to 1 (Equation 1). Under this rule any mean reading time under 3.75 seconds gives a behavioural weight of zero.

$$s = \min(5, 1 + 4t/15), \quad B = \frac{\lfloor s \rfloor - 1}{4}$$

**Equation 1**
*Behavioural weight from the mean reading time t; the brackets denote the whole-number part.*

Second, the two ratings were converted to a *survey weight* S: each rating r was rescaled to 0 to 1, and S was the mean of the rescaled second rating and one minus the rescaled effort rating, so that a type rated easy and, on the system's reading of the second statement, effective received a high weight (Equation 2).

$$S = \frac{r'_2 + (1 - r'_1)}{2}, \quad r' = \frac{r - 1}{4}$$

**Equation 2**
*Survey weight from the effort rating r1 and the second rating r2, each on a 1 to 5 scale.*

Third, the *final weight* was 0.6B + 0.4S, and the six final weights were divided by their sum so that they added to one (Equation 3). The normalised weights W were fixed for the participant throughout the practice sessions.

$$W^* = 0.6B + 0.4S, \quad W = \frac{W^*}{\sum W^*}$$

**Equation 3**
*Final and normalised preference weights.*

**Table 5**
*Worked example: Participant A's preference weights*

| Feedback type | Mean time (s) | Score 1 + 4t/15 | Whole part | Behav. B | Effort | Second | Survey S | Final 0.6B + 0.4S | Norm. W |
|---|---|---|---|---|---|---|---|---|---|
| Elicitation | 11.0 | 3.93 | 3 | 0.50 | 1 | 3 | 0.750 | 0.600 | 0.255 |
| Error code | 9.0 | 3.40 | 3 | 0.50 | 2 | 3 | 0.625 | 0.550 | 0.234 |
| Direct correction | 5.3 | 2.42 | 2 | 0.25 | 2 | 3 | 0.625 | 0.400 | 0.170 |
| Brief explanation | 5.0 | 2.33 | 2 | 0.25 | 1 | 1 | 0.500 | 0.350 | 0.149 |
| Reformulation | 1.3 | 1.36 | 1 | 0.00 | 1 | 2 | 0.625 | 0.250 | 0.106 |
| Location | 3.0 | 1.80 | 1 | 0.00 | 3 | 3 | 0.500 | 0.200 | 0.085 |

*Note.* Participant A is unusual in having six different weights.

### 3.5 Practice Sessions

The application had a React front end and a FastAPI back end; participants used it in a browser on their own devices, most often in the evening. Sessions were unlocked one at a time by the researcher, ran from 12 November to 12 December 2025, a median of 1.1 days apart, and lasted a median of 13 minutes. The item screen (Figure 3A) showed the sentence, four options and a countdown; each item was open for 15 seconds, the chosen option was inserted into the gap and could be changed until the window closed, and unselected items were recorded as unanswered (a median of 19 of 320 per completer). Sessions that were abandoned were voided and restarted (152 attempts).

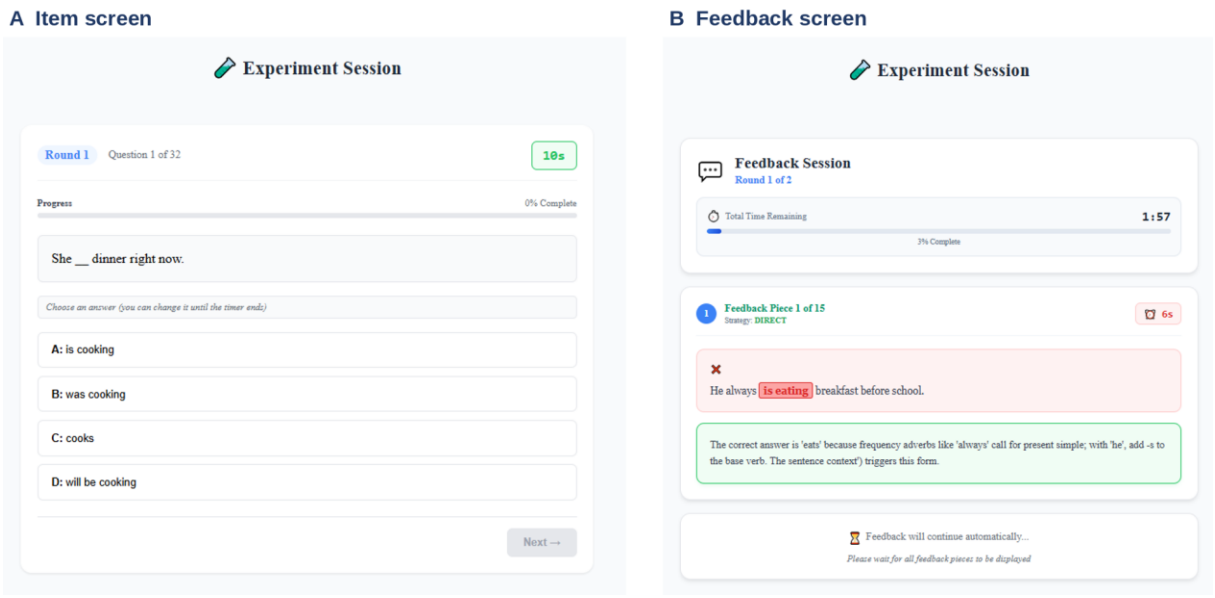


**Figure 3**

*The item screen (A) and the feedback screen (B). In B, the participant's wrong answer is highlighted in the stem and the message appears below it; the counter shows the seconds remaining for this message.*

### 3.6 The Feedback Phase, in Which Time Is the Currency

Three rules governed the feedback phase in both conditions. Every error received exactly one message. The messages shared a time allowance set only by the number of errors: 30 seconds for fewer than five, then 60, 90, 120, 150 and 180 for each further band of five (median 90 seconds among completers). And the allowance was divided among the types present in proportion to the participant's normalised weights, each type's share split equally among its messages, which were then shown one after another for those durations, grouped by type in the GA condition and in item order in the random condition (Figure 3B). Message length played no part. The consequence is the point on which the design turns: the type assigned to an error decided not only what the participant read about it but how long they had to read it. Figure 4 shows this for one participant in the GA condition, referred to below as Participant A and followed through every stage in Sections 3.4.2–3.8.1, in the second round of their first session: eleven errors and 90 seconds; the two one-token error codes were on screen for 10 seconds each, while three elicitation messages of 61 to 78 words received 7 or 8 seconds, which would require reading at 470 to 670 words per minute.

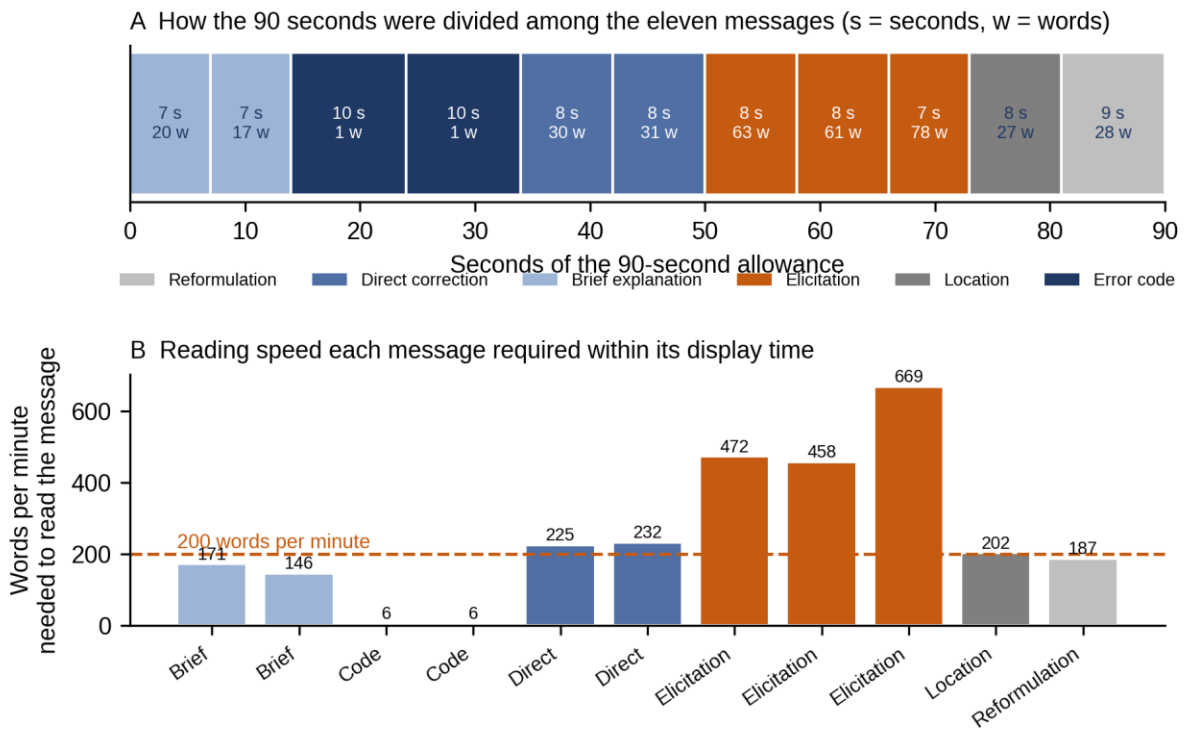


**Figure 4**

*Participant A, session 1, round 2: how the 90-second allowance was divided among eleven messages (A) and the reading speed each message required (B).*

### 3.6.1 Worked Feedback-Allocation Example

$$T_{\text{type}} = \frac{T_{\text{round}} \cdot W_{\text{type}}}{\sum W'}, \quad T_{\text{message}} = \frac{T_{\text{type}}}{n_{\text{type}}}$$

**Equation 4**

*Seconds for a feedback type and for each of its messages; T_round is the time allowance, n_type the number of messages of that type in the round. Shares are rounded to whole seconds so that they sum to the allowance.*

Participant A made eleven errors in the second round of their first session, fixing the allowance at 90 seconds. The algorithm assigned three errors to elicitation, two each to error code, direct correction and brief explanation, and one each to location and reformulation. Their normalised weight for elicitation was 0.255, so elicitation received 23 seconds, split among its three messages as 8, 8 and 7 seconds; error code received 20 seconds, or 10 per message; and so on down to location, which received 8 seconds for its single message. The two error codes, one token each, were on screen for 10 seconds; the three elicitation messages, of 63, 61 and 78 words, were on screen for 8, 8 and 7 seconds, which would require reading at between 470 and 670 words per minute. The whole set of 357 words was allotted 90 seconds. Table 6 gives the full allocation.

**Table 6**

*Worked example: the feedback phase of one round for Participant A (11 errors, 90 seconds)*

| Feedback type | Norm. W | Count | Seconds for the type (W × 90) | Seconds per message | Words per message | Words per minute required |
|---|---|---|---|---|---|---|
| Elicitation | 0.255 | 3 | 23 | 8, 8, 7 | 63, 61, 78 | 470 to 670 |
| Error code | 0.234 | 2 | 20 | 10, 10 | 1, 1 | 6 |
| Direct correction | 0.170 | 2 | 16 | 8, 8 | 30, 31 | 225 to 230 |
| Brief explanation | 0.149 | 2 | 14 | 7, 7 | 20, 17 | 145 to 170 |
| Reformulation | 0.106 | 1 | 9 | 9 | 28 | 185 |
| Location | 0.085 | 1 | 8 | 8 | 27 | 200 |

| Feedback type | Norm. W | Count | Seconds for the type (W × 90) | Seconds per message | Words per message | Words per minute required |
|---|---|---|---|---|---|---|
| Total | 1.000 | 11 | 90 | 90 | 357 | 238 on average |

### 3.7 How the Algorithm Assigned the Types

In the random condition a type was drawn for each error with equal probability. In the GA condition the same rule was used for the first round of the first session, to seed the value estimates, and the algorithm from then on. It scored a candidate assignment by summing, over the errors, the value estimate for the error's tense cell and candidate type times the participant's normalised weight for that type, with repeated uses of a type discounted by successive halving (Equation 5). Because a second use of the highest-weight type earned less than a first use of the next, the algorithm spread errors across types in proportion to the weights: in Participant A's round in Figure 4, three errors went to elicitation, two each to code, direct correction and brief explanation, and one each to location and reformulation, and no assignment could have scored higher. The search used a population of 80, 60 generations, tournament selection, single-point crossover and mutation, with the four best candidates carried forward; it typically stopped improving by generation 6; the full parameter specification is given in Section 3.7.2.

$$\text{Score}(A) = \sum_{i \in \text{errors}} V(type_i, cell_i) \cdot W(type_i) \cdot 0.5^{k_i}$$

**Equation 5**

*Score of a candidate assignment A; k_i is the rank, from zero, of error i among the errors in A that share its type.*

#### 3.7.1 Algorithm Arithmetic and Search Details

Using Equation 5, Participant A's round in Table 6 illustrates the arithmetic. It was their first GA round, so every value estimate was 0.5, and each type's product V × W was simply half its weight: 0.128 for elicitation, 0.117 for error code, 0.085 for direct correction, 0.074 for brief explanation, 0.053 for reformulation and 0.043 for location. Assigning elicitation to a first error earned 0.128; assigning it to a second earned 0.064 and to a third 0.032. Assigning error code to a first error earned 0.117, more than a second elicitation, which is why the algorithm spread the eleven errors across all six types rather than giving them all to the highest-weight type: the assignment in Table 6 scored 0.734, and no assignment of those eleven errors could score higher. The general behaviour follows from this: the algorithm uses the highest-weight types most, moves to the next type once the halving makes it more rewarding, and, when the value estimates are all equal, produces a spread of types that mirrors the weights.

The search itself followed the standard procedure. The algorithm began with 80 candidate assignments generated at random, scored them, carried the four best into the next generation unchanged, and produced the rest by tournament selection, single-point crossover and mutation; it ran for 60 generations and returned the best assignment found. In practice the best score stopped improving after a median of six generations. The random number generator was seeded from the participant, session and round identifiers, so that every plan can be regenerated exactly. The parameters are collected in Section 3.7.2. In both conditions, once the types had been assigned, the time allowance was divided as described in Section 3.6.1.

#### 3.7.2 Specification of the Genetic Algorithm

The table below records the parameters of the algorithm as deployed, recovered from the parameter record stored with every feedback plan.

**Table 7**

*GA parameters as deployed*

| Parameter | Value as deployed |
|---|---|
| Representation of a candidate | One feedback type (index 0 to 5) per wrong item in the round |
| Population size | 80 |
| Number of generations | 60 (best score typically stopped improving by generation 6) |
| Crossover rate | 0.8 |
| Mutation rate | 0.15 |
| Elitism | 5% |

| Parameter | Value as deployed |
|---|---|
| Selection method | Tournament selection, tournament size 3; single-point crossover with probability 0.8; mutation with probability 0.15 per offspring reassigning one error to a random type; the top 5% (4 candidates) carried over unchanged |
| Fitness function | Sum over errors of V(type, cell) × W(type) × $0.5^k$, where k is the number of earlier errors in the candidate already given that type |
| Initial value estimates V | 0.5 for all 72 type-by-cell combinations, per learner |
| Update rule for V | Once per session, comparing round 1 (e1 errors in the cell) with round 2 (e2): delta = r × min(1, e1/3) × (e1 - e2)/e1, applied to every type delivered for that cell in round 1; requires e1 > 0; not bounded; no updates between sessions. As deployed, r = 0.5 in 600 of 617 session-plans and 1.0 in 17. Raw counts are compared although 25 of the 60 session-by-cell combinations have unequal item numbers in the two rounds. |
| Survey weight | (rating of "It takes my tutor a long time to deliver this feedback" normalised to 0 to 1, used as the effectiveness term, + (1 - rating of "This feedback style requires significant mental effort" normalised to 0 to 1)) / 2 |
| Behavioural weight | Mean placement reading time for the type converted by 1 + 4 × (seconds / 15), capped at 5; the whole-number part rescaled by (score - 1) / 4; zero for any mean under 3.75 s |
| Final weight W | 0.6 × behavioural + 0.4 × survey; the six values normalised to sum to 1; fixed for the learner after placement |
| Total feedback time per round | 30 s per five errors, rounded up; minimum 30 s, maximum 180 s |
| Display time per message | Total time × W(type) among the types present, divided equally among that type's messages, rounded to whole seconds by largest remainder; independent of message length; messages shown grouped by type in the GA condition |
| Random condition | One type per error drawn uniformly from the six; same time rule; messages shown in item order; V logged but unused. Also used for the first round of session 1 in the GA condition, to seed the value estimates |
| Random seed | Derived deterministically from the learner, session and round identifiers and stored with the plan |

### 3.8 How the Value Estimates Changed

The value estimates were the only part of the system that changed during practice, once per session. For every tense cell in which the participant had erred in the first round, the system compared the error counts in that cell in the two rounds (e1, e2) and moved the estimate of every type shown for that cell in the first round by Equation 6, upward when errors fell and downward when they rose. As deployed, the rate r was 0.5 in 600 of the 617 session-plans that carried an update and 1.0 in the remaining 17; values were unbounded, and the rule compared raw counts although in 25 of the 60 session-by-cell combinations the two rounds offered different numbers of items, so that a change in the count could reflect the schedule rather than the learner (Section 4.6).

$$\Delta V \;=\; r \;\cdot\; \min\!\left(1, \frac{e_1}{3}\right) \;\cdot\; \frac{e_1 \;-\; e_2}{e_1}, \quad r \;=\; 0.5$$

**Equation 6**

*Change in a value estimate after a session; e1 and e2 are the error counts in the cell in the two rounds; r was 0.5 (1.0 in 17 of 617 session-plans).*

#### 3.8.1 Worked Value-Update Example

Using Equation 6, Participant A's first session shows how the update worked. In the present perfect cell they made three errors in the first round and none in the second; the three types shown for that cell in the first round (brief explanation, elicitation, direct correction) each earned 0.5 × 1 × 3 / 3 = +0.5, and their estimates rose from 0.5 to 1.0. In the present perfect continuous cell they made two errors in the first round and four in the second; the two types shown for it (reformulation, error code) each received 0.5 × (2 / 3) × (2 - 4) / 2 = -0.333, and fell from 0.5 to 0.167. In their next session these estimates made elicitation, direct correction and brief explanation more attractive for present perfect errors, and reformulation and error code less attractive for present perfect continuous errors, than they would otherwise have been.

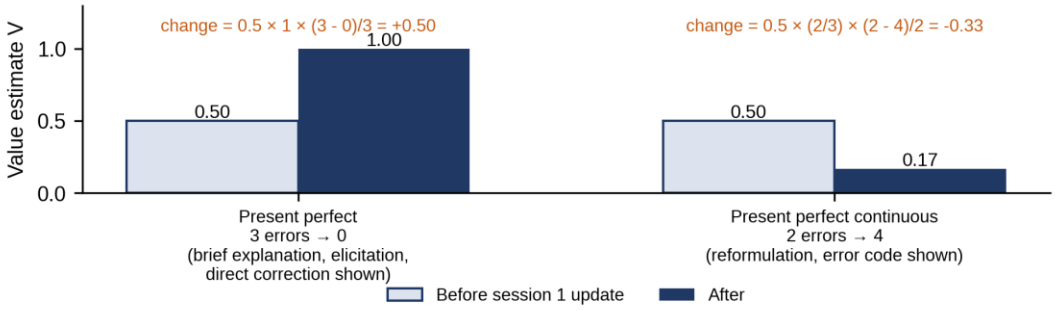


**Figure 5**
*Participant A's value estimates for two tense cells before and after the update at the end of session 1.*

### 3.9 Measures and Analysis

The primary outcome was the correctness of each item response. The protocol excluded responses under 700 ms as too fast to reflect reading; the field concerned records the interval from the first selection of an option to submission rather than time on the item, so the exclusion removed 404 responses; the result is also reported with no exclusion and with all 8,510 sub-700 ms responses excluded. The check on H1 used the concentration of types within a round's messages, measured by the Herfindahl-Hirschman index (HHI), the sum of the squared shares of each type; it runs from 0.167 when all six types are used equally to 1 when one type is used, and takes 0.26 for sets of this size drawn evenly at random. Delivery data for RQ2 were the display time and word count of every message, the placement reading times, the survey ratings and the derived weights.

Concentration of types within a round was measured with the Herfindahl-Hirschman index (Hirschman, 1964; Equation 7).

$$\mathrm{HHI} = \sum_{s=1}^{6} {p_s}^2$$

**Equation 7**
*Concentration of feedback types in a round; p_s is the share of the round's messages that are of type s.*

Accuracy was analysed with a logistic regression fitted by generalised estimating equations (GEE) with participants as clusters, an exchangeable working correlation and robust standard errors (Liang & Zeger, 1986): a logistic regression for correct versus incorrect that allows for each participant contributing about 300 responses. The main model included condition, session (Session 1 as reference) and centred placement proportion; a second added the condition-by-session interaction. Effects are odds ratios (ORs) with 95% confidence intervals (CIs): an OR compares the odds of a correct answer in the random condition with those in the GA condition, 1 means no difference, and the CI gives the range of ORs the data are compatible with. The main model was fitted to the 103 completers; it was also repeated on all 153 starters, with an item-fixed-effects specification, and as a learner-level intention-to-treat approximation with multiple imputation over all 255 allocated participants. The H1 check compared learner-level means between conditions with Welch's t-test, which respects the fact that a learner's rounds are not independent. Given the standard error of the condition effect (0.148 on the log-odds scale), the design could detect an OR of 1.52 or its reciprocal 0.66 with 80% power, so the primary result is interpreted through its interval and an equivalence test against ORs of 0.80 and 1.25 (Lakens, 2017). Analyses were run in Python 3.12 with statsmodels 0.15. The model is specified below, and sensitivity analyses are reported in Section 4.3.1.

$$\log\left[\frac{P(\text{correct})}{1 - P(\text{correct})}\right] = \beta_0 + \beta_1\,\mathrm{Random} + \sum_{k=2}^{5} \beta_k\,\mathrm{Session}_k + \beta_6\,\mathrm{Placement}_c$$

**Equation 8**
*The main model. Random is 1 for the random condition and 0 for the GA condition; Session_k is 1 in session k; Placement_c is the placement proportion minus its mean.*

### 3.10 Transparency and Data

A protocol and analysis plan were written before the outcome analyses but not publicly preregistered; confirmatory analyses are those specified in the protocol. An earlier deployment of the tool with a different cohort is reported in the author's unpublished MSc dissertation (2025); the present analyses supersede it. The de-identified data, item bank, analysis code and application source are available at https://osf.io/nauxj/overview?view_only=cc1c88297d6e49aaba0de3b74dd5b70c . Data handling followed Pardo and Siemens (2014).

## 4. Results

Section 4.1 describes who took part. Section 4.2 checks H1, without which the comparison could not be interpreted. Sections 4.3 to 4.5 answer RQ1 and test H2. Section 4.6 turns to RQ2.

### 4.1 Participant Flow and Baseline

Of the 683 who consented, 328 completed the placement test, 255 were allocated (GA 127, random 128), 153 answered at least one practice item (GA 79, random 74), and 103 completed all five sessions (GA 51, random 52; Figure 6). Retention from allocation to completion was 40% in both conditions ($\chi^2(1) = 0.01$, $p = .94$), and completers had only slightly higher placement scores than non-completers (0.433 versus 0.401, $p = .13$), equally in both conditions; attrition was mildly selective on ability but not differentially so, and its consequences are taken up in Section 5.5. Among completers the conditions did not differ at baseline (GA $M = 0.426$, $SD = 0.180$; random $M = 0.441$, $SD = 0.147$; $p = .64$). Table 8 summarises the completion sample, the basis of all analyses that follow.

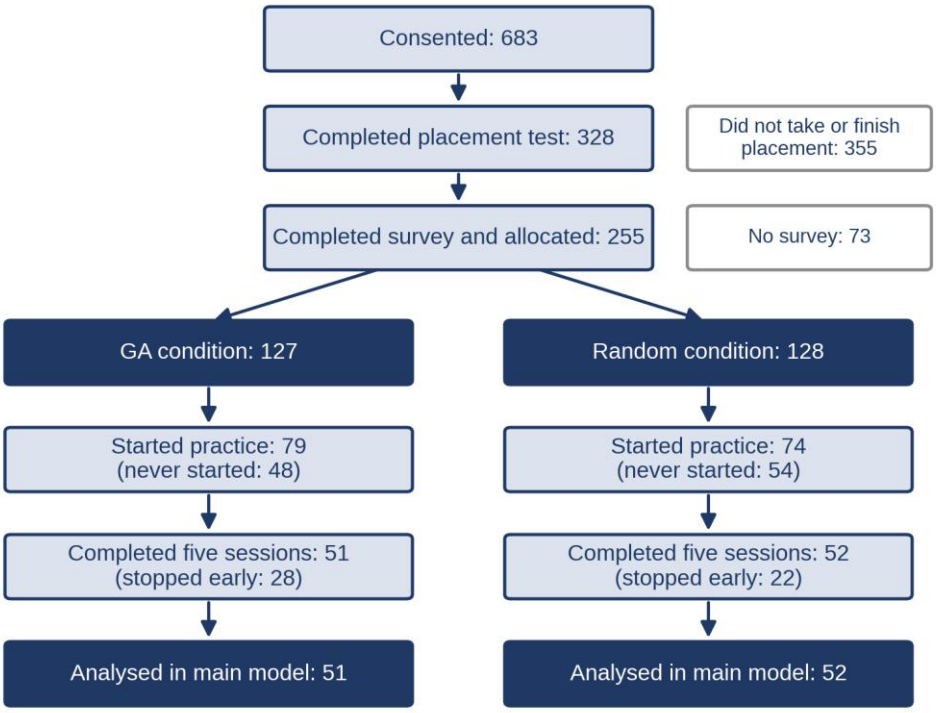


**Figure 6**
*Participant flow from consent to analysis.*

**Table 8**
*Sample characteristics and feedback exposure, completion sample*

| Characteristic | GA (n = 51) | Random (n = 52) | All (n = 103) |
|---|---|---|---|
| Baseline placement proportion, mean (SD) | 0.426 (0.180) | 0.441 (0.147) | 0.433 (0.163) |
| Analysed item responses per learner, median | 296 | 297 | 296 |
| Raw accuracy, % correct | 60.2 | 57.4 | 58.8 |

| Characteristic | GA (n = 51) | Random (n = 52) | All (n = 103) |
|---|---|---|---|
| Feedback rounds delivered | 508 | 522 | 1,030 |
| Time allowance per round in seconds, mean (SD) | 89.5 (36.5) | 91.8 (39.5) | 90.7 (38.0) |
| Messages per round, mean (SD) | 12.0 (5.9) | 12.3 (6.4) | 12.2 (6.2) |
| Messages delivered in total | 6,100 | 6,436 | 12,536 |

### 4.2 H1: Did the Algorithm Change the Feedback Learners Received?

The conditions were matched on the amount of feedback, as the design guaranteed: the time allowance per round had a median of 90 seconds in both, messages per round were similar (medians 11 and 12), and exposure to tense cells was identical. They differed in composition, which is what the algorithm was built to change. Averaging the concentration index over each learner's rounds, GA learners had a mean HHI of 0.215 (SD 0.050) and random learners 0.264 (SD 0.073), a difference of -0.049 (95% CI [-0.073, -0.025], $p < .001$), and GA learners received more distinct types per round (5.33 against 4.92, $p = .016$; Figure 7). H1 is therefore supported. Its size should be kept in view: the random condition sat exactly at the value expected from even random draws (0.26), the GA moved learners about 0.05 towards the floor of 0.167 on a scale that runs to 1, and it left the overall share of each type unchanged, every type accounting for 15% to 18% of messages in both conditions ($\chi^2(5) = 10.36$, $p = .066$). The algorithm spread the six types more evenly within each round, and did nothing else that was measured.

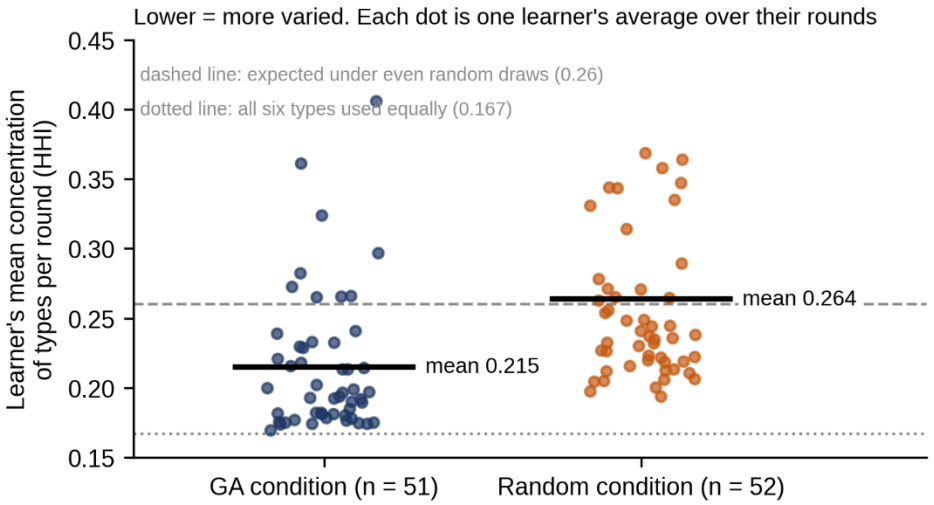


**Figure 7**
*Concentration of feedback types per round (HHI), one dot per learner, with condition means. Lower is more varied.*

**Table 9**
*Feedback messages by type and condition, completion sample*

| Type | GA n | GA % | Random n | Random % |
|---|---|---|---|---|
| Reformulation | 1,061 | 17.4 | 1,006 | 15.6 |
| Direct correction | 1,024 | 16.8 | 1,070 | 16.6 |
| Brief explanation | 1,001 | 16.4 | 1,093 | 17.0 |
| Elicitation | 1,043 | 17.1 | 1,159 | 18.0 |
| Location | 1,053 | 17.3 | 1,078 | 16.7 |
| Error code | 918 | 15.0 | 1,030 | 16.0 |
| Total | 6,100 | 100 | 6,436 | 100 |

*Note.* Counts are of messages actually displayed to completers; 19 planned rounds (about 2%) were never displayed because the session was voided before the feedback phase.

### 4.3 RQ1: Accuracy by Condition

Table 10 reports the main model and Figure 8 the condition effect under it and four alternatives. Averaged across sessions and adjusted for baseline, learners in the random condition had 0.90 times the odds of a correct response of learners in the GA condition (95% CI [0.67, 1.20], p = .48): 57.4% against 60.2% correct, a difference well within what chance would produce with 103 learners. The interval is wide, compatible with GA odds from 17% lower to 49% higher than random. An equivalence test against a smallest effect of interest of 25% in the odds, the shaded band in Figure 8, rejects a random advantage of that size (p = .013) but not a GA advantage (p = .21). The estimate was the same on all 153 starters (OR 0.88 [0.69, 1.12]), under either treatment of fast responses, with a fixed effect for every item (OR 0.87 [0.63, 1.22]), and at learner level with imputation over all 255 allocated participants (random condition 3.0 points lower, 95% CI [-9.0, 2.9]). The findings therefore do not indicate that the two rules were equivalent; they indicate that no advantage was detected, and that a moderate GA advantage in particular remains open.

**Table 10**

*GEE logistic regression predicting item correctness, completion sample (n = 103; 29,823 responses)*

| Predictor | OR | 95% CI | p |
|---|---|---|---|
| Condition: random vs GA | 0.90 | [0.67, 1.20] | .476 |
| Session 2 vs 1 | 0.87 | [0.79, 0.96] | .006 |
| Session 3 vs 1 | 0.97 | [0.88, 1.07] | .508 |
| Session 4 vs 1 | 0.96 | [0.87, 1.06] | .411 |
| Session 5 vs 1 | 0.97 | [0.88, 1.07] | .513 |
| Baseline placement proportion (centred) | 1.41 | [0.54, 3.65] | .481 |

*Note.* OR = odds ratio; Session 1 is the reference. An OR of 0.87 for Session 2 means the odds of a correct answer were 13% lower than in Session 1; a CI that includes 1 means the data are compatible with no difference. Within-learner correlation = 0.13.

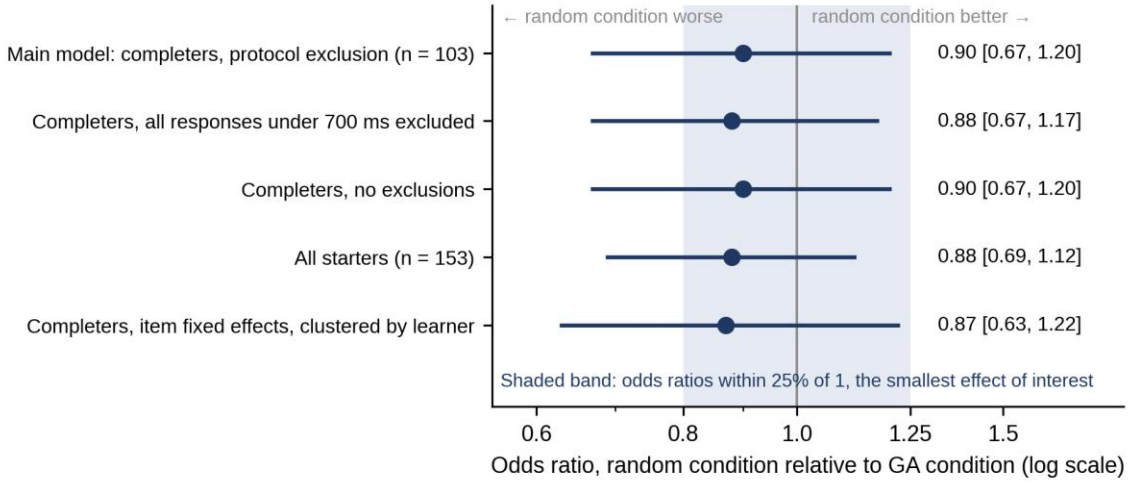


**Figure 8**

*The condition effect under the main model and four alternative treatments of the data. Each dot is an odds ratio and each line its 95% confidence interval; the shaded band marks differences of less than 25% in the odds.*

#### 4.3.1 Sensitivity Analyses

The condition effect was stable across every reasonable choice of sample and exclusion rule (Table 11). Repeating the main model on all 153 participants who contributed analysable responses gave OR = 0.88 (95% CI [0.69, 1.12], p = .30). Excluding all 8,510 responses with recorded times under 700 ms, rather than only the 404 that carried the flag, gave OR = 0.88 (95% CI [0.67, 1.17], p = .38), and excluding nothing gave OR = 0.90 (95% CI [0.67, 1.20], p = .47). A model with a separate effect for every one of the 320 items, which holds item difficulty constant exactly, gave OR = 0.87 (95% CI [0.63, 1.22], p = .43). The learner-level intention-to-treat approximation, which imputed mean accuracy for the 102 allocated participants who never started, estimated the random condition at 3.0 percentage points below the GA condition (95% CI [-9.0, 2.9], p = .32), compared with 3.2 points

among starters and 2.7 points among completers. None of these analyses changes the conclusion, although the last rests on the assumption that non-starters would have performed like starters of the same placement score.

**Table 11**

*Condition effect under alternative samples and exclusion rules*

| Analysis | Learners | Responses | Estimate (random vs GA) | 95% CI | p |
|---|---|---|---|---|---|
| Main model: completers, flagged fast responses excluded | 103 | 29,823 | OR 0.90 | [0.67, 1.20] | .48 |
| Completers, all responses under 700 ms excluded | 103 | 22,736 | OR 0.88 | [0.67, 1.17] | .38 |
| Completers, no exclusions | 103 | 30,177 | OR 0.90 | [0.67, 1.20] | .47 |
| All starters, flagged fast responses excluded | 153 | 34,952 | OR 0.88 | [0.69, 1.12] | .30 |
| Completers, item fixed effects, clustered by learner | 103 | 29,823 | OR 0.87 | [0.63, 1.22] | .43 |
| Learner-level, completers (mean accuracy) | 103 | | -2.7 points | [-9.8, 4.4] | .45 |
| Learner-level, all starters | 153 | | -3.2 points | [-9.1, 2.7] | .28 |
| Learner-level, multiple imputation over all allocated | 255 | | -3.0 points | [-9.0, 2.9] | .32 |

### 4.4 RQ1 and H2: Change Across Sessions

Accuracy did not rise over the five sessions in either condition. Relative to Session 1 the odds of a correct answer were lower in Session 2 (OR 0.87, p = .006) and then about level (ORs 0.96 to 0.97); in raw terms the GA condition moved from 62.5% to 60.9% and the random condition from 57.3% to 57.6% (Figure 9), and the conditions did not follow different trajectories (Wald $\chi^2(4)$ = 3.05, p = .55). H2 was not supported in either part. One caveat applies: each session used a different, non-equated set of 64 items, and a tenth of Session 1 items had been met in placement, so the coefficients describe accuracy on successive item sets rather than learning on a fixed measure. What can be said is that accuracy on the items presented did not rise, and that from Session 3 onwards an adaptive rule had no improvement to amplify.

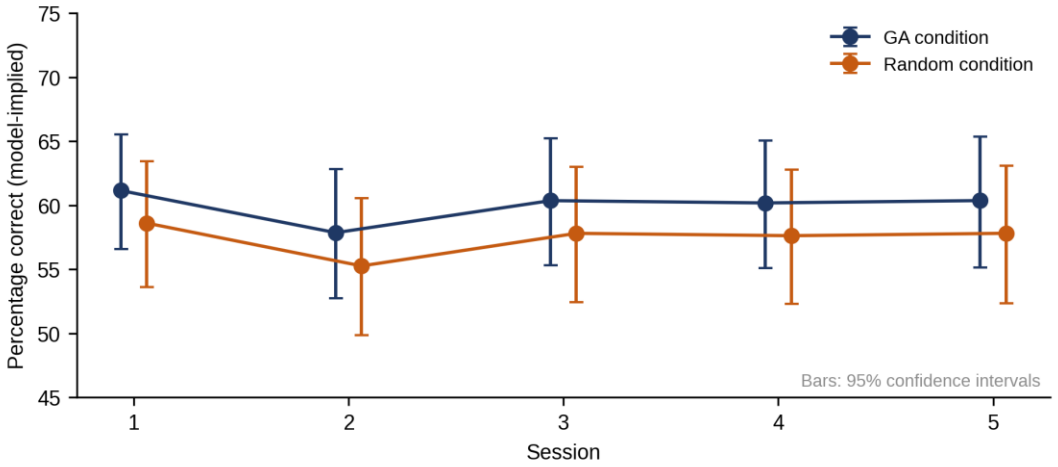


**Figure 9**

*Model-implied percentage correct by condition and session, with 95% confidence intervals.*

### 4.5 Accuracy by Tense Category

Figure 10 gives accuracy by tense cell. Simple forms were easiest (68.5%) and perfect forms hardest (50.2%); past reference was hardest within every aspect, and the extremes were future simple at 79.9% and past perfect continuous at 34.8%, above the 25% chance level for four options. The condition effect did not vary by category (Wald $\chi^2(3)$ = 1.05, p = .79): the GA condition was 2 to 4 points ahead in every category, never reliably so.

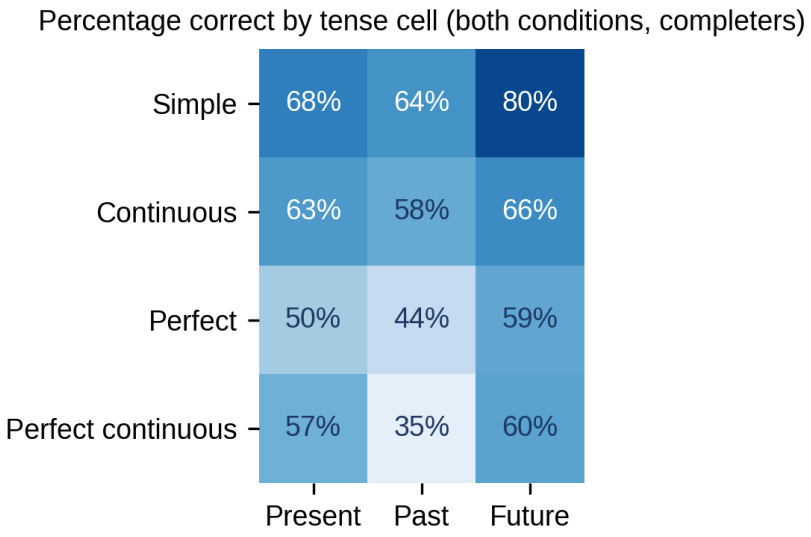


**Figure 10**
*Percentage correct by tense cell, completion sample. Darker cells are easier.*

**Table 12**
*Percentage correct by tense category, time reference and condition, completion sample*

| Tense category | Present | Past | Future | All items | GA | Random |
|---|---|---|---|---|---|---|
| Simple | 68.1 | 64.1 | 79.9 | 68.5 | 69.6 | 67.4 |
| Continuous | 62.5 | 58.4 | 65.5 | 62.1 | 63.7 | 60.5 |
| Perfect | 49.6 | 43.7 | 59.3 | 50.2 | 51.1 | 49.2 |
| Perfect continuous | 56.9 | 34.8 | 60.2 | 54.2 | 56.1 | 52.2 |

*Note.* Each learner answered a median of 73 to 75 items per category. Across sessions, perfect continuous items improved from 50.0% to 57.4% and perfect items declined from 53.7% to 46.8%; the other two categories were flat after Session 1.

### 4.6 RQ2: What the Delivery Data Show

*The composition change was small.* As Section 4.2 showed, the GA produced a measurable but modest change in variety, not a different kind of feedback experience.

*Nearly half of the messages required unusually fast reading to be completed in the time shown.* The median displayed message was 23 words long and on screen for 7 seconds (interquartile range 5 to 9), requiring 192 words per minute; 46% of messages required more than 200 words per minute, and the burden varied by type because time followed weight and not length (Figure 11): a median of 130 words per minute for brief explanations, 188 for reformulations, 210 for location cues, 233 for direct corrections and 531 for elicitation messages, 95% of which exceeded 200. Second-language readers typically read well below the 200 to 300 words per minute of fluent first-language reading (Grabe, 2009); the threshold is a benchmark, not a claim about any individual. For 46% of what was delivered, and nearly all of the most explicit type, the display would have closed before a reader at 200 words per minute finished, in both conditions.

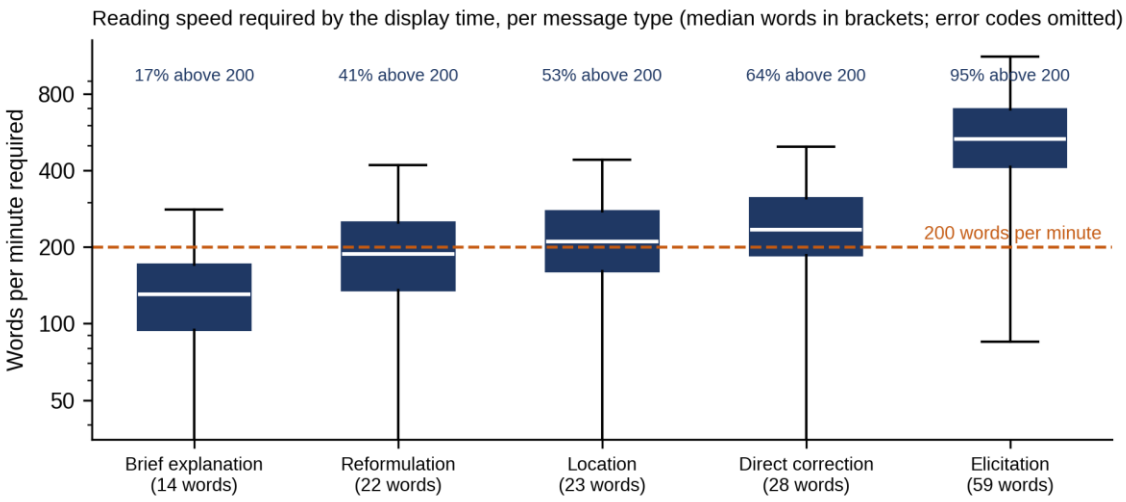


**Figure 11**

*Reading speed required to finish a message within its display time, by type (all messages delivered to completers). Boxes show the middle half of messages.*

*The preference weights carried almost no information.* The reading times from which they were derived were short: across the 328 learners who completed the placement, the median time per message was 2.0 seconds (interquartile range 1.3 to 3.3), 22% of learner-by-type means were at or under one second, and the median was 2 seconds for every one of the six types. Under the rule in Section 3.4.3, a mean under 3.75 seconds gives a behavioural weight of zero; this was the case for 79% of learner-by-type combinations, a further 14% received 0.25, and only 1% reached the maximum. Participant A, with six different weights, was atypical. On the survey side, ratings of the effort statement averaged between 3.2 and 3.4 on the five-point scale for every type, and ratings of the second statement between 3.1 and 3.2, so learners did not differentiate the six types on either. Survey weights were unrelated to behavioural weights across learners (Spearman $r_s$ = -.06, p = .02) and within learners (median $r_s$ = .00, interquartile range -.32 to .36), and effort ratings were unrelated to reading time ($r_s$ = -.03, p = .26). Taken together, these findings indicate that the weights were close to constant (Figure 12): for 46 of the 255 allocated learners all six final weights were identical, and for the median learner the normalised weights ranged only from 0.11 to 0.24 across the six types. For such learners the score in Section 3.7.1 reduced to the value estimate discounted for repetition, and the time each message received depended on little more than how many messages of its type were in the round.

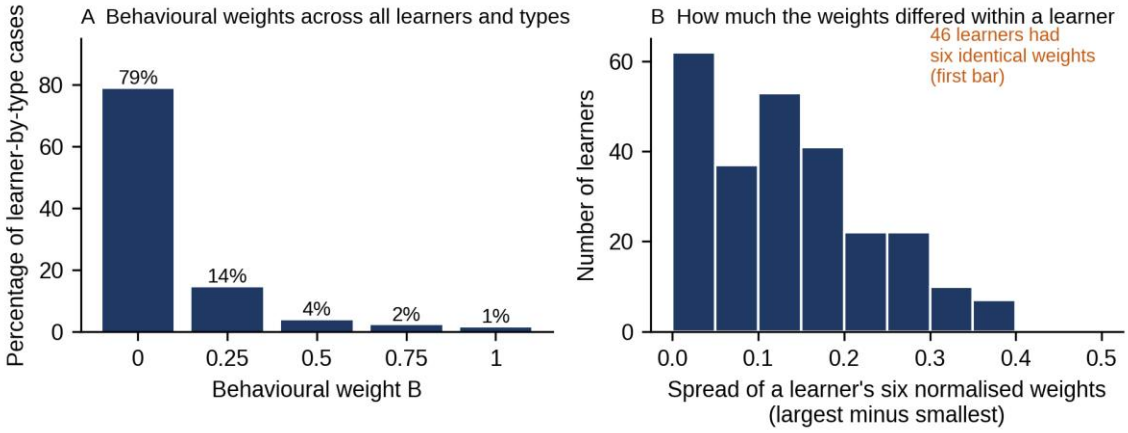


**Figure 12**

*The preference weights across all allocated learners: (A) the behavioural weight was zero in four cases out of five; (B) the spread between a learner's largest and smallest normalised weight.*

*The value estimates were sparse and followed the item schedule.* Estimates were created for a median of 27 of a learner's 72 combinations and a quarter never moved from 0.5. Those that moved were driven in part by the item bank: because the rule compared raw error counts, a cell with fewer items in round 2 tended to produce an upward update and a cell with more items a downward one. Across the 60 session-by-cell combinations, the mean net change in the estimates correlated at rho = -0.84 with the difference in item numbers between the rounds (Figure 13), and among 2,425 learner-session-cell comparisons in which the

error count changed, its direction disagreed with the direction of the error rate per item in 9% of cases overall and 22% where the item numbers differed. The adaptive component had, for most learners and cells, little to adapt with, and part of what it adapted to was the design of the bank.

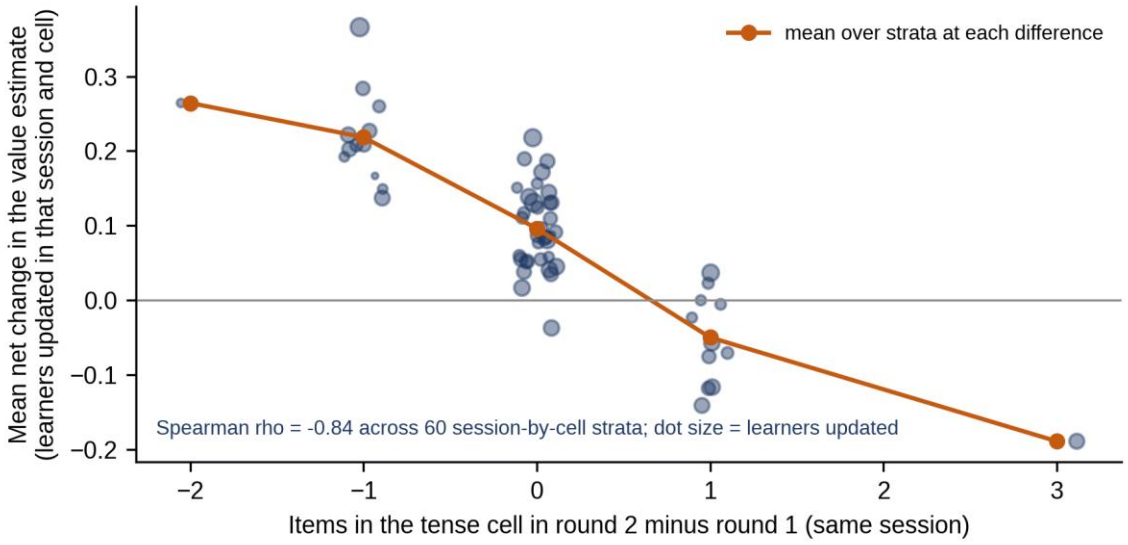


**Figure 13**

*How the value updates followed the item schedule. Each dot is one session-by-tense-cell combination: its height is the mean net change in the value estimates of learners updated for that cell, its position the number of items in round 2 minus round 1.*

## 5. Discussion

### 5.1 Summary of Findings

This study examined what a controlled evaluation can establish about an algorithmic rule for assigning corrective feedback types, and what the tool's own logs add. With regard to RQ1, the design did what it was built to do: the check confirmed that the algorithm changed the feedback sets, and the comparison then found no detectable difference in accuracy between the two conditions, with a point estimate slightly favouring the GA and an interval too wide to rule out a moderate advantage; the conclusion held under every alternative analysis and in every tense category, and neither condition improved across five sessions. With regard to RQ2, the delivery data indicate that the change in composition was small, that 46% of messages required reading above 200 words per minute to be completed in time, and that the learner-specific inputs to the rule were close to constant or followed the item schedule.

### 5.2 Why Did a Working Algorithm Produce No Advantage?

Five features of the study account for the result, in descending order of likely importance.

First, there was no upward accuracy signal to amplify. Accuracy on the items presented was flat from Session 1 to Session 5 in both conditions, so a rule could only have shown an advantage by producing improvement where the observed session profile showed none. Feedback arrived in a batch after 32 items, minutes after the error and on another screen; learners could not retry the item or answer the elicitation question; and the items required recognition rather than production. The CF literature ties the effect of feedback to immediate engagement with it, as uptake in interaction (Lyster & Ranta, 1997) or revision in writing (Bitchener & Ferris, 2012), and the tutorial CALL evidence comes from systems in which feedback is attached to the learner's response and followed by another attempt (Heift, 2004, 2010; Nagata, 1993). Feedback after multiple-choice testing improves later retention (Butler & Roediger, 2008), but the outcome here was accuracy on the next round of different items. The format resembled a series of tests with a debrief more than tutoring.

Second, time was allocated by preference rather than by need. A message's time followed the learner's weight for its type, not its length, so a one-token code received the same seconds as a 63-word elicitation message and 46% of messages required more than 200 words per minute to be read in full. Messages that cannot reasonably be read in full have little opportunity to differ in their effects, whatever rule chose them; this affected both conditions equally and is consistent with the null. It also reverses the intended logic of the weights: giving a type the learner had read longer more time is defensible, but giving a long message of a low-weight type less time than a short message of a high-weight type is not.

Third, the inputs to the score carried little information. The value estimates were created only for cells in which the learner erred, updated once per session, and moved with the item schedule almost as much as with the learner. The preference weights were zero on their behavioural component in four fifths of cases, came from ratings that did not distinguish the six types, and were identical across types for a fifth of learners. An optimiser that maximises the product of a sparse value and a near-constant weight produces assignments close to arbitrary, which is what the type shares in Section 4.2 show.

Fourth, the manipulation was small: random draws from six types already produce varied sets, and the GA moved concentration by 0.05 on a scale whose floor is 0.167 without changing the overall share of any type. If the effect of type is itself modest under text-only conditions, as the meta-analyses suggest, a modest redistribution within rounds could not be expected to register in a sample of this size. Fifth, the outcome was immediate accuracy on multiple-choice items; gains in production, delayed retention or transfer were not measured.

### 5.3 What the Evaluation Design Contributes

The main contribution is methodological, and it can be stated as four checks that an evaluation of adaptive feedback should pass in order, each possible from the same logs: whether the rule changed the feedback learners received; whether the feedback could be read in the time given; whether the learner-specific inputs to the rule carried information about the learner; and only then whether accuracy differed. This study passed the first modestly, failed the second and third in ways the logs make plain, and found no difference at the fourth. Read in that order, the null is diagnostic rather than merely disappointing. Two consequences follow. An evaluation needs an active baseline with matched exposure, including matched time, since a comparison with less feedback would be uninformative about the rule; the random condition is demanding precisely because it holds the intended exposure variables constant while changing the rule. And the rule is downstream of the format and its delivery: a single-condition pilot showing improvement across sessions, with a check that messages can be read and that the inputs vary across learners and are comparable from round to round, would have been a cheaper and more informative first step than a two-condition comparison. Studies in this area often report only the outcome comparison (Doroudi et al., 2019); the present result shows why that is not enough.

### 5.4 Pedagogical Implications

Variety of feedback type is not in itself a benefit. What appears to matter is whether the learner engages with the feedback, not how many kinds there are.

Display time must follow message length. A rule of thumb for adult second-language readers is at least half a second per word, more for metalinguistic content; here the longest type was shown for the same seven seconds as the shortest.

Feedback should stay close to the error: shown immediately at the item, on the same screen, with a second attempt. An elicitation prompt that cannot be answered, or that answers itself, is a prompt in name only, and an internal code without a key is not feedback.

Learners' ratings of feedback types are not a guide to their behaviour, and three two-second reading times are not a basis for personalisation. Designers should check that a personalisation input varies across learners before building on it.

In the absence of evidence favouring adaptive assignment, explicit correction with a brief rule is a defensible default (Heift, 2004; Nagata, 1993; Van der Kleij et al., 2015); and voluntary out-of-class practice on personal devices, which retained 15% of those who consented, needs to be built into coursework if it is to reach most learners.

### 5.5 Limitations

Attrition was severe: 60% of allocated participants did not complete. Retention did not differ by condition and the completer, starter and imputation analyses agreed, but the imputation rests on untestable assumptions and selective attrition can distort balanced comparisons (Zhou & Fishbach, 2016). The design could detect only an odds ratio of about 1.5. Allocation was balanced alternation rather than randomisation, unbiased only if approval order was unrelated to participants' characteristics, which cannot be verified. The three ingredients of the algorithm were confounded, so no one of them can be assessed alone. Four features of the deployed system made the adaptive inputs less informative than intended: the second survey statement concerned delivery time and was used as the effectiveness term; the behavioural weight used the whole-number part of the preference score; the value estimates moved at a rate of 0.5, and of 1.0 in 17 session-plans; and the update compared raw counts across rounds with unequal item numbers. None favoured either condition. Messages were shown grouped by type in the GA condition and in item order in the random condition, an unintended difference confounded with the rule. The placement drew from the practice bank; the response-time field measures choice-to-submission latency, not time on the item; sessions used non-equated item sets; the outcome was immediate recognition accuracy; no demographic or experience data were collected; and the study was conducted at one institution with one language background and was not preregistered.

### 5.6 Directions for Future Research

A follow-up should fix delivery and format before the algorithm: messages displayed for a time that scales with their length or until dismissed, with dismissal logged; feedback at the item with a retry, on production items; a separate placement pool; and a single-condition pilot showing improvement before rules are compared. Comparisons should include a value-only rule, a repetition-penalty-only rule and a fixed "correction plus rule" default. Value estimates should be updated every round and compare error rates per item rather than raw counts; preference weights, if kept, need many more exposures, items that measure what the objective assumes, and a check that they vary across learners. Outcomes should include a delayed post-test and a production task, attrition should be reduced by embedding practice in a course, and the study should be preregistered with an intention-to-treat analysis and a smallest effect of interest fixed in advance (Lakens, 2017).

## 6. Conclusion

In conclusion, this study has examined the assignment of corrective feedback types in a browser-based English tense tutor, with a specific focus on what a controlled comparison against an exposure-matched active baseline can establish about a genetic algorithm that weighs learner-specific value, preference and variety, and on what the tool's own delivery logs add to that. The findings indicate that the algorithm can be embedded in a CALL tool, can assign feedback types according to a stated objective, and can be shown through logged data to change the composition of the feedback learners receive, but that it did not improve accuracy relative to random assignment, and that no upward session-wise accuracy trend was observed in either condition. The logs and the source code suggest why: the fixed time allowance was divided by preference rather than by need, so that much of the feedback required reading rates above a reasonable benchmark to be completed in time, and the learner-specific inputs to the algorithm were too sparse and too nearly constant to carry information, while the one input that did change followed the item schedule. While the study does not claim to provide an exhaustive account of adaptive feedback selection, it offers a design, with an active baseline, matched exposure, a check on what the algorithm changed, and an audit of how the feedback was delivered, that separates the question of whether an algorithm altered the intervention from the question of whether the altered intervention improved learning. It therefore contributes to CALL research by showing that changing the mix of feedback is not, by itself, enough, and that an algorithm for choosing among feedback messages can be no better than the practice format and the delivery conditions within which it operates.

### Statements and Declarations

Funding: This research received no external funding. No external funding has been received for the article processing charge.

Conflicts of Interest: The author reports no competing interests.

### Data, Code and Materials Availability

The de-identified dataset, codebook, analysis code, item bank and application source code are available at https://osf.io/nauxj/overview?view_only=cc1c88297d6e49aaba0de3b74dd5b70c. The reproduction script regenerates the reported results and worked examples from the data package.

### References

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use.* Blackwell.

Bitchener, J., & Ferris, D. (2012). *Written corrective feedback in second language acquisition and writing.* Routledge. https://doi.org/10.4324/9780203832400

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56-64). Worth Publishers.

Boubekri, A., & Ech-Charfi, A. (2023). A contrastive analysis of aspect in English and Moroccan Arabic. *Lingua Posnaniensis, 65*(1), 25-47. https://doi.org/10.14746/linpo.2023.65.1.2

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616. https://doi.org/10.3758/MC.36.3.604

Clement, B., Roy, D., Oudeyer, P.-Y., & Lopes, M. (2015). Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining, 7*(2), 20-48.

Doroudi, S., Aleven, V., & Brunskill, E. (2019). Where's the reward? A review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education, 29*(4), 568-620. https://doi.org/10.1007/s40593-019-00187-x

El Hadari, T. (2022). An optimality-theoretic account of the acquisition of English tenses: The case of Moroccan EFL learners. *Southern African Linguistics and Applied Language Studies, 40*(3), 306-324. https://doi.org/10.2989/16073614.2022.2063146

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge University Press. https://doi.org/10.1017/CBO9781139150484

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Heift, T. (2004). Corrective feedback and learner uptake in CALL. *ReCALL, 16*(2), 416-431. https://doi.org/10.1017/S0958344004001120

Heift, T. (2010). Prompting in CALL: A longitudinal study of learner uptake. *The Modern Language Journal, 94*(2), 198-216. https://doi.org/10.1111/j.1540-4781.2010.01017.x

Heift, T., & Schulze, M. (2007). *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues.* Routledge.

Heift, T., & Schulze, M. (2015). Tutorial computer-assisted language learning. *Language Teaching, 48*(4), 471-490. https://doi.org/10.1017/S0261444815000245

Hirschman, A. O. (1964). The paternity of an index. *The American Economic Review, 54*(5), 761-762.

Holland, J. H. (1975). *Adaptation in natural and artificial systems.* University of Michigan Press.

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

Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal, 99*(1), 1-18. https://doi.org/10.1111/modl.12189

Kartal, G., & Yeşilyurt, Y. E. (2024). A bibliometric analysis of artificial intelligence in L2 teaching and applied linguistics between 1995 and 2022. *ReCALL, 36*(3), 359-375. https://doi.org/10.1017/S0958344024000077

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284. https://doi.org/10.1037/0033-2909.119.2.254

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*(2), 309-365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Liang, J.-C., Hwang, G.-J., Chen, M.-R. A., & Darmawansah, D. (2023). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments, 31*(7), 4270-4296. https://doi.org/10.1080/10494820.2021.1958348

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22. https://doi.org/10.1093/biomet/73.1.13

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition, 19*(1), 37-66. https://doi.org/10.1017/S0272263197001034

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition, 32*(2), 265-302. https://doi.org/10.1017/S0272263109990520

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching, 46*(1), 1-40. https://doi.org/10.1017/S0261444812000365

Mitchell, M. (1998). *An introduction to genetic algorithms.* MIT Press.

Nagata, N. (1993). Intelligent computer feedback for second language instruction. *The Modern Language Journal, 77*(3), 330-339. https://doi.org/10.1111/j.1540-4781.1993.tb01980.x

Nassaji, H., & Kartchava, E. (Eds.). (2021). *The Cambridge handbook of corrective feedback in second language learning and teaching.* Cambridge University Press. https://doi.org/10.1017/9781108589789

Paas, F. G. W. C., & van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors, 35*(4), 737-743. https://doi.org/10.1177/001872089303500412

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology, 45*(3), 438-450. https://doi.org/10.1111/bjet.12152

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning, 31*(7), 653-674. https://doi.org/10.1080/09588221.2018.1428994

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133-164). John Benjamins. https://doi.org/10.1075/lllt.13.09val

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207-217. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c332. https://doi.org/10.1136/bmj.c332

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Tan, L. Y., Hu, S., Yeo, D. J., & Cheong, K. H. (2025). Artificial intelligence-enabled adaptive learning platforms: A review. *Computers and Education: Artificial Intelligence, 9*, 100429. https://doi.org/10.1016/j.caeai.2025.100429

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327-369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475-511. https://doi.org/10.3102/0034654314564881

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197-221. https://doi.org/10.1080/00461520.2011.611369

Verschoor, A. J. (2007). *Genetic algorithms for automated test assembly* [Doctoral dissertation, University of Twente].

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19-38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111*(4), 493-504. https://doi.org/10.1037/pspa0000056