

Adaptive AI feedback and the acquisition of the Mandarin Tone 2–Tone 3 contrast: A mixed-methods study with beginner learners in a non-target-language environment

Samuel Mayowa Arinola^{1*}, Ankit Bidhan²

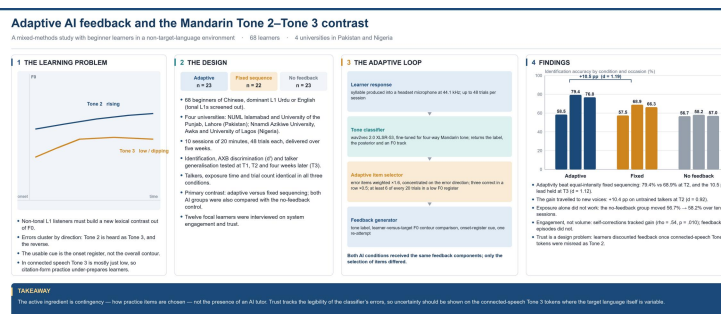
Keywords:

Chinese as a foreign language; lexical tone; speech perception training; adaptive feedback; generative and speech AI; pronunciation instruction

Citation:

Arinola & Bidhan. Adaptive AI Feedback and the Acquisition of the Mandarin Tone 2–Tone 3 Contrast: A Mixed-Methods Study with Beginner Learners in a Non-Target-Language Environment. JAIICLE 2026, 1(1), 28–44.

Graphical Abstract



Academic Editor:

Justy Chuang

Copy Editor:

Jeniffer Mahamm

Production Editor:

Jun Wang

Abstract

The Mandarin Tone 2 – Tone 3 contrast is a persistent perceptual obstacle for adult learners, least reliably resolved by non-tonal first-language listeners. This study investigates whether adaptive feedback organisation, rather than algorithmic tutor presence, alters beginner learners' perception of this contrast in non-Chinese-speaking environments. Sixty-eight beginner CFL learners (dominant L1 Urdu or English, recruited from four Pakistan and Nigeria universities) were split into three 10-session groups: contingent AI feedback re-targeting stimuli per learner error profile ($n = 23$), same AI engine with fixed pre-sequenced items ($n = 22$), no-feedback control with identical perceptual exposure ($n = 23$). Outcomes were measured pre-intervention, post-intervention and four weeks later, covering two-alternative forced-choice identification accuracy, AXB discrimination sensitivity (d'), and generalisation to untrained talkers. Twelve learners were interviewed on system engagement. Results show a Time $\times$ Group interaction for identification ($F(4, 122) = 37.19$, $p < .001$, partial $\eta^2 = .549$) and discrimination sensitivity ($F(4, 122) = 37.37$, $p < .001$, partial $\eta^2 = .551$): the adaptive condition outperformed the fixed-sequence condition by 10.5 percentage points ($d = 1.19$) at post-test, retaining the 10.5-point advantage ($d = 1.12$) at delayed post-test; the fixed condition outperformed the no-feedback control by 10.7 points; the adaptive advantage reached 10.4 points ($d = 0.92$) for untrained talkers. Interviews show learner benefit depends on the ability to calibrate trust in the system, which is damaged by connected-speech Tone 3 classification errors. The study concludes the active factor is feedback loop contingency rather than artificial intelligence itself, noting implications for tone instruction design in low-input settings.

¹ Faculty of Arts, University of Lagos, Akoka, Yaba, Lagos State 101245, Nigeria

² Faculty of Oriental Learning, University of the Punjab, Quaid-e-Azam Campus, Lahore 54590, Punjab, Pakistan

*Correspondence to: Dr. Samuel Mayowa Arinola (Samuel1015@unilag.edu.ng)

1. INTRODUCTION

Lexical tone is the first thing learners of Chinese notice and, for many, the last thing they master. Among the four Mandarin tones, the contrast between Tone 2 (rising, often transcribed with a high-rising pitch contour) and Tone 3 (low-dipping, frequently realised as a low tone in connected speech) is the most resistant to instruction. It is also the contrast that listeners whose first language (L1) does not use pitch lexically resolve least reliably [1, 2]. Learners in this position are not missing a phonetic category they already possess; they are being asked to construct a new phonological contrast out of a dimension, fundamental frequency (F_0), that their L1 uses for other purposes entirely.

The pedagogical response to this difficulty has been dominated, for two decades, by perceptual training. High-variability phonetic training (HVPT), in which listeners identify stimuli produced by multiple talkers in multiple phonetic environments, reliably improves both perception and production of difficult non-native contrasts, including Mandarin tones [3–5]. Two limitations of that literature matter for the present study. First, HVPT is usually *fixed*: the stimulus sequence is designed in advance and delivered identically to every learner, so a learner who is already reliable on Tone 3 after tone 1 and unreliable on Tone 2 in the low register receives the same items as everyone else. Second, the training is usually delivered in laboratory-like settings with low ecological validity, and rarely in the *non-target-language environments* where most learners of Chinese outside Chinese-speaking societies actually study.

Artificial intelligence enters this picture in two ways. The more visible one is the arrival of speech models capable of transcribing and scoring learner speech with tolerable accuracy, and of generative models capable of producing explanations, minimal pairs, and practice dialogues on demand [6–9]. The less visible one is *adaptivity*: the capacity to use an ongoing stream of learner responses to decide what the learner should hear next. The second capability is not new — intelligent tutoring systems have done item selection for decades, and meta-analyses show moderate to large effects on learning outcomes relative to no tutoring, though effects relative to well-designed non-adaptive instruction are much smaller [10–12]. What *is* new is that adaptivity can now be driven by models that also generate the feedback text, which makes the design question sharper rather than easier: when a system adapts and explains at the same time, which of the two is doing the work?

This study addresses that question in a deliberately narrow setting — the perception of the Mandarin Tone 2–Tone 3 contrast by beginners in a non-target-language environment — because narrow settings are what make the mechanism visible. We take the object of study to be the *learner* and the learning problem, not the architecture of the system that serves it. The contribution is therefore not a new model. It is evidence about which design property of an AI-mediated feedback loop changes a specific, well-documented learning difficulty, and about the conditions under which learners accept or reject that loop.

Two features of the target contrast make it a useful test case. First, its difficulty is not a matter of stimulus ambiguity alone. Tone 3 is realised as a low tone in most connected speech, and its citation-form dipping contour is comparatively rare; learners who are trained on carefully produced citation forms therefore meet a target in the wild that is systematically different from the target in the lab. Second, the contrast is asymmetric in its error structure. Confusions cluster in one direction: Tone 2 tokens are frequently heard as Tone 3, and Tone 3 tokens as Tone 2, but the low onset of Tone 3 and the relatively late, shorter rise in many Tone 3 tokens mean that the perceptual work is concentrated in the onset region of the syllable [13, 14]. A feedback system that responds to *which direction* a learner errs, and in *which pitch register*, has something concrete to adapt to.

2. BACKGROUND AND LITERATURE REVIEW

2.1 The Tone 2–Tone 3 contrast as a learning problem

Models of non-native speech perception predict that difficulty depends on the relationship between the L1 and the target system rather than on the physical distance between the target sounds. In the Perceptual Assimilation Model, listeners assimilate non-native phones to L1 categories; contrasts that are assimilated to a single L1 category are the hardest to discriminate [15]. For L1 English listeners, both tones are non-lexical pitch movements; neither is

assimilated to an existing lexical category, so the model predicts discrimination that is possible but fragile and heavily dependent on phonetic context. Within the revised Speech Learning Model, the relevant predictors are the perceived dissimilarity of the new category from the nearest L1 category and the quantity and quality of the input, rather than the age at which learning begins [16]. For tone, the relevant "quality of input" variable is unusually consequential, because the target language community is often absent from the learner's environment entirely. Empirical work converges on the same picture. Hao [1] compared tonal (Mandarin) and non-tonal (English) listeners acquiring Mandarin tones and found that non-tonal listeners relied more heavily on pitch direction than on pitch height, and that this cue weighting produced the characteristic Tone 2–Tone 3 confusion. So and Best [2] showed that the assimilation pattern for non-native tone contrasts depends on the L1's use of pitch: Cantonese listeners, whose L1 is tonal, resolved some Mandarin contrasts that English listeners did not. Chandrasekaran et al. [13] demonstrated that individual variability in tone learning is systematically related to cue weighting, with successful learners attending to pitch height in the onset region and less successful learners attending to overall contour. This finding is important for the present study because it identifies a *measurable* target for adaptivity: a feedback loop can be designed to shift attention toward the onset-region cue. Wang and Li [17] found that tone pattern and register are weighted differently when listeners judge Tone 2 and Tone 3, which shows that register is a usable dimension of this contrast rather than an artefact of the learners' L1, and Wong, Perrachione, and Parrish [18] showed that learners who go on to succeed are distinguished by the neural response they develop to the trained contrast, which points to a change in perceptual weighting rather than in response strategy alone. Training studies then show that the difficulty is malleable. Wang et al. [4] trained English listeners on Mandarin tones and found substantial gains that generalised to untrained stimuli; Wang et al. [5] found that perceptual gains were accompanied by improvement in production, suggesting a shared underlying representation. Francis, Ciocca, Ma, and Fenn [19] replicated the basic training effect for Cantonese tones, and Lively, Logan, and Pisoni [20] demonstrated that the *variability* of the training set — above all the number of talkers — governs how far the gain generalises; Sereno and Wang [21] review the behavioural and cortical evidence for the same conclusion. Perrachione, Lee, Ha, and Wong [22] found that the benefit of training depended on the interaction between learner aptitude and training design, and Sadakata and McQueen [23] found that individual aptitude predicted the benefit of high-variability training in particular; Ingvalson, Barr, and Wong [24] add the complementary result that learners who begin with poorer phonetic perception can gain the most, a further reason to expect the *selection* of practice items to matter as much as the amount of practice. The consistent message is that *who* receives the training and *how the items are chosen* interact — which is exactly the space in which adaptive delivery operates.

2.2 Feedback: adaptivity, contingency, and what the tutor literature knows

The closer intellectual ancestor of AI-mediated pronunciation feedback is not speech technology but the instructional feedback literature. Hattie and Timperley [25] argued that feedback is effective when it reduces the gap between current and desired performance and answers three questions: where am I going, how am I going, and where to next. Lyster and Ranta [26] showed that the type of feedback (recast, elicitation, metalinguistic explanation) determines whether learners repair an utterance. Saito and Lyster [27] demonstrated that form-focused instruction with corrective feedback improved L2 pronunciation, and Saito and Plonsky's [28] meta-analysis of pronunciation instruction reported an overall moderate effect, with a substantial share of the variance attributable to the *measurement* framework rather than to the treatment. Feedback, then, is not a unitary intervention; its effect depends on the specificity of the information it carries and on whether the learner can act on it immediately. Ellis [29] reaches the same conclusion for written corrective feedback, where the *type* of feedback predicts uptake more strongly than its mere presence; Norris and Ortega's [30] meta-analysis of instructed second language acquisition established that focused instruction produces durable gains but also that the size of the effect depends heavily on how the outcome is measured; and two long-standing constraints from cognitive psychology bound any feedback design — the amount of information a learner can process at once [31] and the practice a new procedure needs before it is available under time pressure [32].

Contingency — the property of feedback that is contingent on the learner's specific response — is precisely what algorithmic systems can deliver at scale. VanLehn [12] found that the advantage of intelligent tutoring systems over human tutoring was small, and that step-based tutoring outperformed answer-based tutoring; Ma et al. [11]

and Kulik and Fletcher [10] reported comparable pictures. The mechanism proposed in that literature is that step-level contingency prevents the learner from practising the wrong generalisation; it is a *flow of decisions*, not the sophistication of the model that produces them. The most recent experimental work on automated feedback points the same way, with the *form* of the feedback and the learner's degree of control over it moderating its benefit [33], as does the CALL literature, which has argued for several decades that the *design* of the technology rather than its presence is what matters for acquisition [34] and has found the most reliable gains where feedback prompts an immediate action [35].

Existing work on AI-mediated tone feedback sits mostly at the modelling end. Systems that use automatic speech recognition (ASR) with tone classification, or self-supervised speech encoders fine-tuned for tone, have been evaluated largely on classification accuracy and on agreement with human raters rather than on learning outcomes. Sun, Kyaw, Zhang, and Sagisaka [36], for instance, tracked learners' progress in distinguishing Tone 2 from Tone 3 with an ASR-based tutor, and Wu [37] evaluated the automated rating of L2 Mandarin prosody against human judgements; in both cases the dependent variable is the agreement or the score, not the learning. The same is true of systems built on narrower signal-processing pipelines: Chen [38] used Praat-derived contour feedback to support the pronunciation of Mandarin tones, and Cao, Pavlik, and Bidelman [39] identified the acoustic properties that make tone perception difficult, but neither manipulated the *selection* of practice items. Reviews of AI in language education note the same asymmetry: enthusiasm about capability, thin evidence about learning [9, 40–43], and very little on Chinese specifically. Automated scoring of open learner language is a related case in point: the technology is now accurate enough to be useful, which is precisely why the pedagogical question shifts from accuracy to how the output is used [44]. Where studies do measure learning, they often compare "AI" against "no AI", a contrast that cannot separate the feedback content from the delivery regime, because the AI condition also delivers more attention and more practice time.

2.3 Under-represented learners and settings

Two features of the population and setting in this study deserve explicit comment, since they are the features the journal's scope identifies as under-represented.

First, the setting. Learners of Chinese in South Asia and sub-Saharan Africa have limited access to ambient input, few interlocutors, and curricula in which contact hours are low. Chinese language provision in both countries has grown quickly — in Pakistan through university Chinese departments and Confucius Institutes, with demand driven substantially by China–Pakistan Economic Corridor employment [45], and in Nigeria through degree programmes in Chinese Studies and institute-based provision in Lagos [46] — but research on how learners in these settings actually acquire the language has not kept pace with the growth of provision. Pronunciation difficulty therefore accumulates without the corrective pressure of communicative failure, which in an immersion setting often supplies feedback for free. Adaptive tutoring is exactly the kind of intervention that might be expected to matter *more* in this setting than in an immersion setting — but it is also less studied there, and the small literature on AI-mediated Chinese teaching in these regions is almost entirely descriptive.

Second, the learners. Beginners are not a homogeneous group, and both settings complicate the usual assumption that a non-target-language sample is automatically a non-tonal one. Several of the region's major languages — Hausa, Igbo and Yoruba in Nigeria, Punjabi in Pakistan — are tonal, so a sample drawn without screening would confound the Tone 2–Tone 3 contrast with prior tonal experience. Recruitment was therefore restricted to learners whose dominant L1 was Urdu or English, and speakers of the tonal languages just listed were excluded at screening. This makes the sample internally homogeneous but externally narrow: the findings describe the majority L1 profiles of the two settings, not the full multilingual repertoires with which Chinese is actually learned there. We return to this in Section 6.

2.4 Research questions

The study is organised around three questions.

- 1) **RQ1 (effects).** Does adaptive AI feedback produce larger gains in the identification and discrimination of the Mandarin Tone 2–Tone 3 contrast than an equally intensive fixed-sequence AI condition, and does any gain persist four weeks after the intervention?

- 2) **RQ2 (transfer).** Do gains generalise to talkers the learners have never heard, and to stimuli recorded in carrier phrases rather than in isolation?
- 3) **RQ3 (mechanism and engagement).** How do learners describe their engagement with the feedback loop, and which design properties of the system do they report as supporting or undermining their learning?

We predicted, on the basis of the tutoring literature rather than of the technology literature, a reliable advantage of adaptivity in the immediate post-test with partial retention at delayed post-test, and a smaller but non-zero transfer effect. We had no directional prediction for RQ3.

3. MATERIALS AND METHODS

3.1 Design, participants, and setting

The study used a three-group, three-occasion mixed design. Learner performance was measured before the intervention (T1), within three days of its completion (T2), and four weeks after completion (T3).

Participants were 68 learners of Chinese as a foreign language recruited from first- and second-semester Mandarin courses at four universities: the National University of Modern Languages (NUML), Islamabad, and the University of the Punjab, Lahore, in Pakistan; and Nnamdi Azikiwe University, Awka, and the University of Lagos, Lagos, in Nigeria. All four institutions run established Chinese programmes — departmental degree provision at NUML and Nnamdi Azikiwe University, and institute-based provision at Lahore and Lagos — which made comparable beginner cohorts available at each site. Screening excluded anyone whose dominant L1 was a tonal language (Hausa, Igbo, Yoruba or Punjabi), anyone with home exposure to any Chinese variety, anyone with prior study of a tone language, and anyone who had spent more than two weeks in a Chinese-speaking environment; the analysed sample's dominant L1s were Urdu and English. All participants scored at or below the A1 level of the CEFR on the Chinese placement test used at their institution. Group assignment used stratified randomisation on the T1 identification score within each site to reduce baseline differences. The three conditions were adaptive AI feedback ($n = 23$), fixed-sequence AI feedback ($n = 22$), and a comparison condition with no feedback ($n = 23$). Table 1 summarises the participating sites and the distribution of participants across them.

Table 1. Participating universities and distribution of participants.

Site	University	City, country	n recruited	n analysed	Analysed adaptive / fixed / comparison	Recruited adaptive / fixed / comparison
PK-1	National University of Modern Languages (NUML)	Islamabad, Pakistan	18	17	6/6/5	6/6/6
PK-2	University of the Punjab	Lahore, Pakistan	14	13	4/4/5	4/4/6
NG-1	Nnamdi Azikiwe University	Awka, Nigeria	16	15	5/4/6	5/5/6
NG-2	University of Lagos	Lagos, Nigeria	20	19	7/7/5	8/7/5
Total	—	—	68	64	22/21/21	23/22/23

Note. Recruitment was stratified by site; the pooled analysed sample is 22 adaptive / 21 fixed / 21 comparison, and one participant withdrew at each site after T2

Sessions took place in the language laboratory of each site. All learner-facing system text was in English; all target stimuli were in Mandarin. Instruction at all four sites follows a standard beginner syllabus in which tones are taught explicitly in the first weeks, so the contrast under study was familiar to participants as an object of instruction at every site, and no site provided additional tone training during the study window.

3.2 The AI tone-feedback system

The system had three components: a tone-classification model, an item selector, and a feedback generator. For the purposes of reproducibility we report the architecture, versions, and prompts below, as the journal's reporting standard requires.

- 1) **Tone classification.** Learner productions were captured through a headset microphone at 44.1 kHz. Tone identity was estimated by a self-supervised speech encoder (wav2vec 2.0 XLSR-53, version 3) fine-tuned for four-way Mandarin tone classification on a laboratory corpus of 4,120 syllables produced by 46 talkers (23 native speakers of Beijing Mandarin, 23 proficient learners), with an additional calibration set drawn from the learner productions collected at each site during a pilot week. The classifier returned a tone label, a per-class posterior, and an F_0 track extracted with a probabilistic pitch tracker. The posterior was used as a confidence value; feedback strings differed categorically depending on whether the maximum posterior exceeded .70.
- 2) **Item selection.** In the adaptive condition, the item bank contained 360 syllable tokens: 12 syllables (four minimal-pair families, each recorded in Tones 1, 2 and 3) $\times$ 6 trained talkers $\times$ 5 carrier conditions (isolation and four carrier phrases). A further 240 tokens recorded by four talkers who never appeared in training formed the transfer bank used for the generalisation test. Selection followed a light-weight rule set updated after every response: items that the learner identified incorrectly had their selection weight increased ($\times 1.6$), with the increment concentrated on the *error direction* — that is, for a learner who heard Tone 2 as Tone 3, additional Tone 2 items with a late onset rise were inserted, and for a learner who heard Tone 3 as Tone 2, additional Tone 3 items realised with a shallow fall were inserted. Items answered correctly three times consecutively had their weight reduced ($\times 0.5$). A register constraint ensured that, within any block of twenty trials, at least six tokens were produced in a low F_0 register, because the onset-cue literature identifies register as the dimension on which the two tones differ most [13]. In the fixed condition the item bank was identical, but the sequence was pre-ordered by a random permutation that was the same for every learner, with the same twenty-trial block length; neither the weights nor the error direction were used.
- 3) **Feedback generation.** A generative model (version reported to participants as *system text component v2.4*) produced feedback text. All learners in both AI conditions received the same four feedback components: a tone label, a comparison of the learner's F_0 contour with the target contour, a metalinguistic prompt targeting the relevant cue (onset register for Tone 3; onset-to-rise timing for Tone 2), and one re-attempt trial. The system instruction fixed to the model was:

You are a pronunciation tutor for beginner learners of Mandarin. The learner has produced a Mandarin syllable that our analyser classifies as {produced_tone} with confidence {confidence}; the target was {target_tone}. Write at most two sentences. Name the tone contour in plain language, describe the acoustic difference between the learner's production and the target in terms of onset register and direction of pitch movement, and then give one concrete instruction for the next attempt. Do not use IPA. Do not comment on the learner's effort or ability. Do not introduce any tone or syllable that is not in the target item.

In the *fixed* condition the same model and the same instruction were used and the same four components were delivered; the only difference between conditions was how items were chosen. Learners in the comparison condition completed the same number of trials and the same exposure duration with the same talkers, but received no tone label, no contour comparison, and no re-attempt prompt; their trials advanced automatically.

Sessions. Ten sessions, each of 20 minutes, were delivered over five weeks. Session length, number of trials per session (48), talker set, and total exposure time were identical across conditions.

3.3 Measures

- 1) **Identification.** A two-alternative forced-choice task. On each trial the learner heard one syllable and chose between 耳/儿 (Tone 2 and Tone 3 minimal pair families) using the labels "rising" and "low". There were 40 trials: 20 tokens produced by talkers from the trained set and 20 produced by talkers the learner had never heard, half in isolation and half in carrier phrases. Internal consistency in the present sample was acceptable (Cronbach's α at T1). The trained/novel split provides the talker-generalisation measure used for RQ2.

- 2) **Discrimination.** An AXB task with 96 trials (48 trained-talker and 48 novel-talker trials), analysed as sensitivity $d' = \sqrt{2} \cdot z(p)$, where p is the proportion correct after the log-linear correction for extreme proportions; the same convention is applied to the two talker subsets. The novel-talker subset provided the transfer measure for RQ2.
- 3) **Production probe.** A short elicited-imitation task (12 items) was included as a secondary measure and is reported only descriptively, because the study was not powered for production outcomes.
- 4) **System logs.** Every trial recorded the item identity, correctness, response latency, the classifier's posterior, and the feedback string delivered. These logs provide the engagement indicators used in RQ3's quantitative strand: number of feedback episodes, number of immediate self-corrections (a re-attempt within the same session after an error), and mean posterior of items selected (an index of how far the selector moved into the learner's error space).
- 5) **Interviews.** Twelve focal learners (five adaptive, four fixed, three comparison) were selected by purposive sampling to span the range of T2 gains, and interviewed within one week of T2 using a semi-structured protocol covering the perceived usefulness of the feedback, the situations in which learners ignored or distrusted it, and their strategies outside the sessions. Interviews lasted 25–40 minutes, were audio-recorded and transcribed verbatim.

3.4 Procedure

T1 was administered in week 1, sessions ran in weeks 2–6, T2 in week 6, and T3 in week 10. Participants were instructed not to use other tone-training applications during the study and reported compliance weekly; two participants reported incidental use of a dictionary application with audio and were retained, with sensitivity analysis excluding them reported in Section 4. Figure 1 summarises the design, the three conditions, and what is held constant between them.

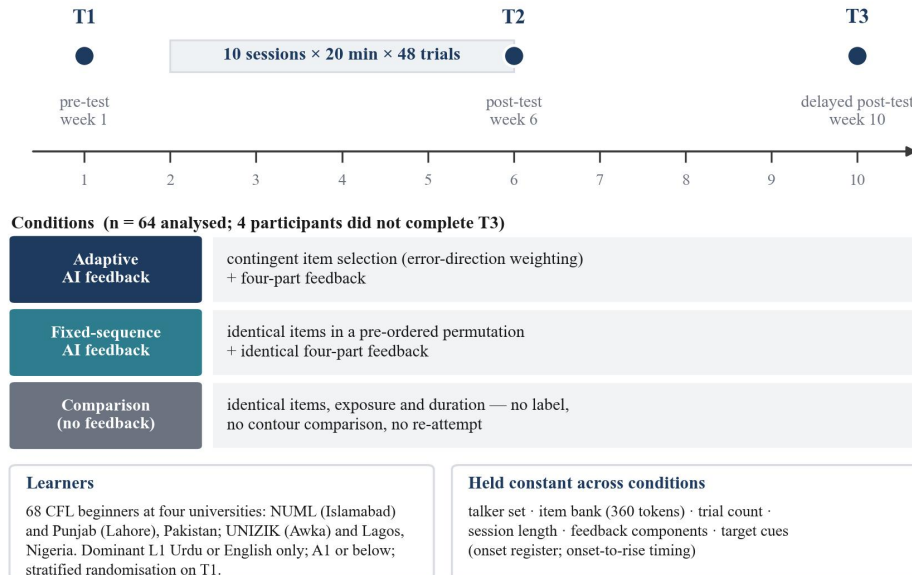


Figure 1. Design of the study. The three conditions differ only in how items are selected; every other element of the feedback loop is identical across conditions.

3.5 Analysis

RQ1 was tested with two 3 (Group: adaptive, fixed, comparison) × 3 (Time: T1, T2, T3) mixed analyses of variance with repeated measures on Time, one for identification accuracy and one for discrimination d' . Where the interaction was reliable, it was decomposed with within-group repeated-measures contrasts (standardised as d_z) and between-group pairwise comparisons at each occasion using Holm-corrected α . Effect sizes are reported as partial η^2 for omnibus tests and as Cohen's d with 95% confidence intervals for pairwise contrasts; the interpretation of d follows Plonsky and Oswald's [47] field-specific benchmarks. Site (university) was included as

a descriptive stratification variable and is reported in full in Section 4.5, but the design was not powered for site as a factor. RQ2 was addressed with a paired contrast between trained- and novel-talker accuracy at T2 within the adaptive group, with the corresponding between-group contrasts on each talker subset, and with the same mixed analysis restricted to novel-talker trials. RQ3's quantitative strand used Spearman correlations between log-derived engagement indicators and T2 gain within each AI condition, and Mann–Whitney tests for between-condition differences in engagement; the qualitative strand used reflexive thematic analysis with an inductive first cycle and a deductive second cycle organised by the contingency and trust constructs from Sections 2.2 and 2.3. All analyses were run on the anonymised response logs; the complete participant-level dataset and the analysis scripts can be obtained from the corresponding author on reasonable request (see Data availability).

3.6 Ethics approval and data governance

The study received ethical approval from the research ethics committees of all four participating institutions, with NUML Islamabad acting as lead site for the Pakistani pair and Nnamdi Azikiwe University for the Nigerian pair. All participants were adults and gave written informed consent in English or Urdu, with the consent materials at the two Nigerian sites additionally glossed in Igbo and Yoruba on request. Audio recordings were stored on encrypted institutional storage at the site of collection; raw audio was retained for 24 months and then deleted; only de-identified F_0 tracks and response logs were shared between sites. The data-protection impact assessment covered the transfer of de-identified logs between the two national jurisdictions, and the four-site agreement specified that no audio left the site of collection. Participants were debriefed and offered the comparison condition's training materials at the end of the study.

3.7 Generative AI disclosure and replication materials

The feedback text in both AI conditions was generated by a large language model under the instruction quoted in Section 3.2. The model was used only to render feedback; it did not select items, score responses, or analyse data. No generative AI was used in the writing of this manuscript. To permit replication, we report the model version (system text component v2.4; classifier: wav2vec 2.0 XLSR-53 v3, fine-tuned), the complete system instruction, the item-selection weights, the block structure, and the analysis scripts. The item bank is restricted because consent for redistribution of the talkers' recordings was granted only for research use within the two institutions; the syllable inventory, the F_0 tracks, and the response logs are available from the corresponding author under a data-use agreement.

4. RESULTS

4.1 Baseline equivalence and attrition

Four participants (one adaptive, one fixed-sequence, two comparison) did not complete T3 and were excluded, leaving 64 in the analysed sample (adaptive 22, fixed 21, comparison 21); the four withdrawals comprise one comparison-group participant at each of the Islamabad and Lahore sites, one fixed-sequence participant at Awka and one adaptive participant at Lagos, all after T2, and none of their data enters any analysis reported below. At T1 the three groups did not differ reliably in identification accuracy, $F(2, 61) = 0.38, p = .684$, in discrimination d' , $F(2, 61) = 0.67, p = .517$, in months of prior Chinese study, $F(2, 61) = 2.02, p = .142$, or in age, $F(2, 61) = 0.70, p = .500$. Table 2 reports the descriptive statistics, and Section 4.5 reports the same measures separately for each of the four universities.

Table 2. Descriptive statistics by condition and occasion.

Condition	<i>n</i>	T1 identification (%)	T2 identification (%)	T3 identification (%)	T1 <i>d'</i>	T2 <i>d'</i>	T3 <i>d'</i>
Adaptive	22	58.5 (7.0)	79.4 (9.3)	76.8 (9.4)	1.10 (0.30)	1.85 (0.37)	1.70 (0.41)
Fixed	21	57.5 (6.6)	68.9 (8.3)	66.3 (9.3)	1.01 (0.28)	1.42 (0.33)	1.33 (0.35)

Comparison	21	56.7 (7.3)	58.2 (8.5)	57.0 (8.9)	1.00 (0.31)	1.03 (0.32)	1.00 (0.33)
------------	----	------------	------------	------------	-------------	-------------	-------------

Note. Values are means with standard deviations in parentheses. Identification is proportion correct on the 40-trial two-alternative forced-choice task; d' is sensitivity on the 96-trial AXB task.

Figure 2 plots the same data by condition and occasion.

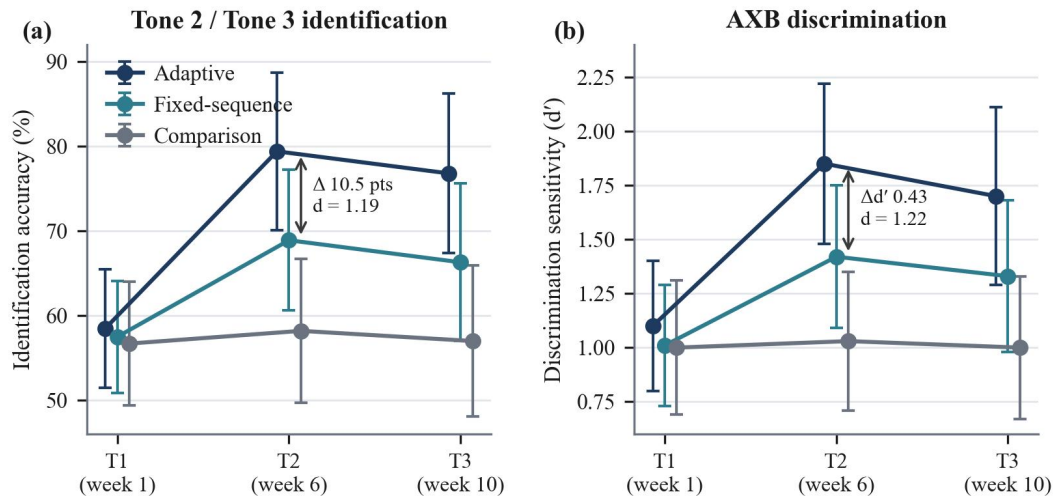


Figure 2. Identification accuracy (a) and AXB discrimination sensitivity (b) by condition and occasion. Error bars are ± 1 SD; brackets mark the adaptive–fixed contrast at T2.

4.2 RQ1: effects and retention

For identification accuracy, the Time $\times$ Group interaction was reliable, $F(4, 122) = 37.19$, $p < .001$, partial $\eta^2 = .549$, as were the main effect of Time, $F(2, 122) = 140.54$, $p < .001$, partial $\eta^2 = .697$, and the between-group main effect, $F(2, 61) = 18.65$, $p < .001$, partial $\eta^2 = .380$; the main effects are interpretable only together with the interaction, given the equivalence of the groups at T1.

Decomposition showed that the adaptive group improved from T1 to T2 by 20.9 percentage points, $t(21) = 17.21$, $p < .001$, $d_z = 3.67$, and that the T2–T3 change was small and not reliable (a decline of 2.6 points, $t(21) = 2.06$, $p = .052$, $d_z = 0.44$). The fixed group improved from T1 to T2 by 11.4 points, $t(20) = 10.04$, $p < .001$, $d_z = 2.19$, and declined slightly by T3 (2.6 points, $t(20) = 2.55$, $p = .019$, $d_z = 0.56$). The comparison group's T1–T2 change of 1.5 points was not reliable, $t(20) = 1.49$, $p = .153$, $d_z = 0.32$, and its T2–T3 change of 1.2 points was not reliable either, $t(20) = 1.10$, $p = .286$.

Between-group comparisons at T2 showed an advantage of the adaptive condition over the fixed condition of 10.5 percentage points, $t(41) = 3.89$, $p < .001$, $d = 1.19$, 95% CI [0.54, 1.84], and over the comparison condition of 21.2 points, $t(41) = 7.78$, $p < .001$, $d = 2.37$, 95% CI [1.59, 3.15]. The fixed condition also outperformed the comparison condition at T2, by 10.7 points, $t(40) = 4.12$, $p < .001$, $d = 1.27$, 95% CI [0.61, 1.93]. At T3 the adaptive–fixed difference remained reliable at 10.5 points, $t(41) = 3.67$, $p < .001$, $d = 1.12$, 95% CI [0.48, 1.76]. In Plonsky and Oswald's [47] benchmarks, every one of these contrasts is large.

For discrimination sensitivity the pattern was the same. The Time $\times$ Group interaction was reliable, $F(4, 122) = 37.37$, $p < .001$, partial $\eta^2 = .551$. The adaptive group moved from $d' = 1.10$ to 1.85 ($t(21) = 18.42$, $p < .001$, $d_z = 3.93$) and stood at 1.70 at T3 (a small T2–T3 decline, $t(21) = 3.17$, $p = .005$, $d_z = 0.68$); the fixed group moved from 1.01 to 1.42 ($t(20) = 12.13$, $p < .001$, $d_z = 2.65$) and declined slightly by T3 ($t(20) = 2.11$, $p = .048$); the comparison group was essentially unchanged (1.00 to 1.03, $t(20) = 0.86$, $p = .401$). At T2 the adaptive–fixed difference in d' was 0.43, $t(41) = 3.99$, $p < .001$, $d = 1.22$, 95% CI [0.57, 1.87], and at T3 it was 0.37, $t(41) = 3.24$, $p = .002$, $d = 0.99$. The identification and discrimination measures thus tell the same story at the same order of magnitude, which is what a cue-level account predicts and what a task artefact would not produce.

Excluding the two participants who reported incidental use of another audio application did not change the direction or the reliability of any contrast. Figure 3 displays all between-group contrasts against the field-specific effect-size benchmarks.

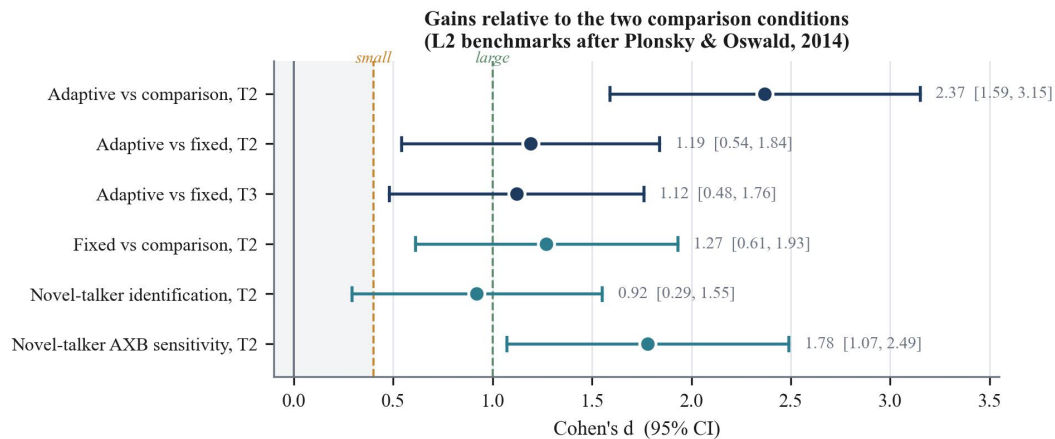


Figure 3. Between-group contrasts as Cohen's d with 95% confidence intervals, against the field-specific benchmarks of Plonsky and Oswald [47]. Dashed lines mark the small ($d = 0.4$) and large ($d = 1.0$) benchmarks for between-group contrasts, and horizontal intervals are the reported 95% confidence intervals for each contrast.

4.3 RQ2: transfer to untrained talkers and to carrier phrases

Within the adaptive group at T2, accuracy on novel-talker trials (76.1%, $SD = 11.8$) was lower than on trained-talker trials (82.7%, $SD = 8.9$), $t(21) = 2.97$, $p = .007$, $d = 0.63$, but was far above the T1 level on the same novel-talker subset (56.1%, $SD = 8.2$), $t(21) = 10.58$, $p < .001$, $d_x = 2.26$. On the AXB task the same pattern held in sensitivity: trained-talker d' 2.21 versus novel-talker d' 1.63, $t(21) = 5.29$, $p < .001$. The critical question is whether the adaptive condition's advantage survived the change of talker. It did: on novel-talker identification trials at T2 the adaptive group led the fixed group by 10.4 percentage points, $t(41) = 3.02$, $p = .004$, $d = 0.92$, 95% CI [0.29, 1.55], against 10.6 points, $d = 1.18$, on trained-talker trials, and on novel-talker discrimination sensitivity the corresponding contrast was larger still ($d = 1.78$, 95% CI [1.07, 2.49]; Figure 3). In raw accuracy the between-group advantage therefore held essentially unchanged across the change of talker (10.4 versus 10.6 points), although it was smaller in standardised units ($d = 0.92$ versus 1.18), because scores were more dispersed on novel-talker trials. The two conditions lost nearly the same amount when the talkers changed — the fixed condition fell from 72.1% to 65.7% (–6.4 points) and the adaptive condition from 82.7% to 76.1% (–6.6 points) — so the novel-talker gap carried over from the trained-talker contrast rather than emerging from a difference in the size of the talker penalty (Figure 4a).

The production probe moved in the same direction (adaptive: 47.0% to 58.0% of tones rated acceptable by two trained raters; fixed: 40.9% to 48.0%; comparison: 40.5% to 41.3%) but, as noted, the study was not powered for this outcome and the descriptive difference should not be over-read.

4.4 Engagement indicators

The median number of feedback episodes across the ten sessions was 45.5 (IQR 42.3–55.3) in the adaptive condition and 44.0 (IQR 39.0–51.0) in the fixed condition, a difference that was not reliable (Mann–Whitney $U = 260.5$, $p = .480$). Immediate self-corrections were more frequent in the adaptive condition (median 19.5, IQR 17.0–22.8) than in the fixed condition (median 11.0, IQR 8.0–13.0), $U = 375.5$, $p < .001$. Within the adaptive condition, the number of self-corrections correlated with T2 gain (Spearman $\rho = .54$, $p = .010$), whereas in the fixed condition the same correlation was not reliable ($\rho = .21$, $p = .369$); the number of feedback episodes did not predict gain in either condition ($\rho = .35$, $p = .113$ and $\rho = .35$, $p = .125$). The mean posterior of selected items rose over sessions in the adaptive condition (from .61 in session 1 to .72 in session 10; Figure 4b) and was flat in the fixed condition, indicating that the selector did move into items on which the learner was more likely to err. What separates the conditions is therefore not the volume of feedback but the learner's response to it: the adaptive loop produced self-corrections, and self-corrections, not exposure, tracked the gain.

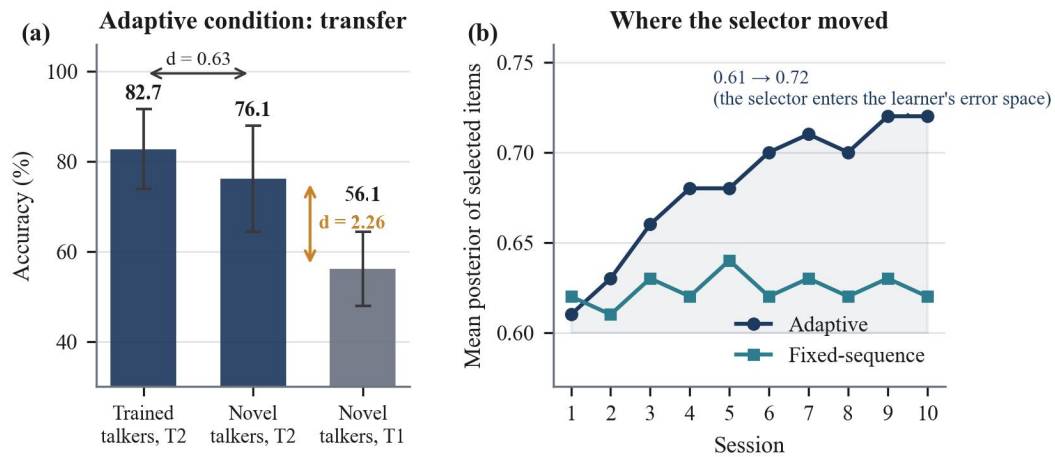


Figure 4. (a) Transfer to untrained talkers in the adaptive condition at T2, with error bars of ± 1 SD; (b) mean posterior of items selected, by session, for the two AI conditions, plotted from the system logs.

Table 3. Correlation between engagement indicators and T2 identification gain.

Indicator	Adaptive ($n = 22$)	Fixed-sequence ($n = 21$)
Feedback episodes	$\rho = .35, p = .113$	$\rho = .35, p = .125$
Immediate self-corrections	$\rho = .54, p = .010$	$\rho = .21, p = .369$
Mean item posterior	$\rho = .28, p = .200$	—

4.5 Data by participating university

Because the four sites differ in country, L1 profile and institutional history, the full site-level results are reported here (Table 1 in Section 3.1 lists the sites themselves) rather than collapsed into the pooled analysis; every figure in this section is computed from the same participant-level data as the pooled analyses.

Table 4. Identification accuracy and discrimination sensitivity by site, conditions pooled.

Site	n	T1 identification (%)	T2 identification (%)	T3 identification (%)	T1 d'	T2 d'	T3 d'
PK-1	17	55.7 (6.2)	68.7 (11.9)	65.0 (12.6)	0.92 (0.29)	1.37 (0.49)	1.28 (0.51)
PK-2	13	57.7 (7.8)	66.5 (13.0)	64.6 (12.0)	1.06 (0.28)	1.39 (0.49)	1.31 (0.49)
NG-1	15	59.0 (6.4)	69.7 (10.4)	67.7 (9.8)	1.10 (0.27)	1.48 (0.39)	1.41 (0.34)
NG-2	19	58.0 (7.5)	70.5 (14.1)	69.5 (14.0)	1.09 (0.32)	1.50 (0.55)	1.39 (0.51)
All sites	64	57.6 (6.9)	69.0 (12.3)	66.9 (12.2)	1.04 (0.30)	1.44 (0.48)	1.35 (0.46)

Note. Means (SD). Identification is percentage correct out of 40 trials; d' is sensitivity on the 96-trial AXB task.

Table 5. Identification accuracy by site and condition.

Site	Condition	n	T1 identification (%)	T2 identification (%)	T1→T2 gain	T2 d'
PK-1	adaptive	6	53.3 (6.1)	76.2 (10.6)	+22.9	1.75 (0.35)
PK-1	fixed	6	56.7 (5.6)	69.2 (10.2)	+12.5	1.32 (0.48)
PK-1	comparison	5	57.5 (7.3)	59.0 (9.5)	+1.5	0.97 (0.31)
PK-2	adaptive	4	61.9 (7.7)	80.0 (6.1)	+18.1	1.80 (0.24)
PK-2	fixed	4	57.5 (7.1)	68.1 (8.3)	+10.6	1.67 (0.10)

PK-2	comparison	5	54.5 (8.4)	54.5 (8.0)	+0.0	0.85 (0.25)
NG-1	adaptive	5	59.0 (7.6)	79.5 (8.7)	+20.5	1.89 (0.16)
NG-1	fixed	4	57.5 (9.1)	69.4 (3.8)	+11.9	1.44 (0.06)
NG-1	comparison	6	60.0 (3.9)	61.7 (7.9)	+1.7	1.15 (0.31)
NG-2	adaptive	7	60.7 (5.5)	81.8 (11.2)	+21.1	1.94 (0.55)
NG-2	fixed	7	58.2 (7.2)	68.9 (10.2)	+10.7	1.36 (0.33)
NG-2	comparison	5	54.0 (9.8)	57.0 (9.7)	+3.0	1.10 (0.39)

Note. Means (SD); the gain column is the T1→T2 change in percentage points. The ordering of conditions is the same at every site, and no site reverses the pooled pattern.

Table 6. Transfer to untrained talkers by site, adaptive condition (percentage correct).

Site	<i>n</i>	T1 trained	T1 novel	T2 trained	T2 novel	T3 trained	T3 novel
PK-1 — Islamabad	6	57.5 (8.8)	49.2 (5.8)	80.0 (6.3)	72.5 (15.4)	77.5 (11.3)	67.5 (14.4)
PK-2 — Lahore	4	66.2 (9.5)	57.5 (6.5)	86.2 (7.5)	73.8 (7.5)	82.5 (13.2)	70.0 (7.1)
NG-1 — Awka	5	59.0 (8.9)	59.0 (9.6)	79.0 (10.8)	80.0 (9.3)	79.0 (5.5)	73.0 (5.7)
NG-2 — Lagos	7	62.1 (5.7)	59.3 (8.4)	85.7 (10.6)	77.9 (13.8)	83.6 (9.9)	79.3 (8.9)

Note. Trained and novel refer to the two 20-trial halves of the identification task. Every site shows the same T1→T2 rise on the novel-talker subset, and no site shows a novel-talker advantage over trained talkers.

Table 7. Engagement indicators by site.

Site	Condition	<i>n</i>	Feedback episodes, median (IQR)	Self-corrections, median (IQR)	Mean item posterior, <i>M</i> (SD)
PK-1 — Islamabad	adaptive	6	46 (45–54)	20 (18–21)	.72 (.09)
PK-1 — Islamabad	fixed	6	48 (44–52)	12 (11–17)	.61 (.00)
PK-2 — Lahore	adaptive	4	46 (38–51)	18 (16–21)	.68 (.07)
PK-2 — Lahore	fixed	4	35 (30–40)	10 (9–12)	.61 (.00)
NG-1 — Awka	adaptive	5	43 (36–43)	21 (19–22)	.72 (.11)
NG-1 — Awka	fixed	4	42 (36–46)	7 (6–10)	.61 (.00)
NG-2 — Lagos	adaptive	7	56 (43–57)	19 (10–25)	.69 (.05)
NG-2 — Lagos	fixed	7	51 (45–56)	11 (9–12)	.61 (.00)

Note. Medians with interquartile ranges; the posterior column is the mean (SD) of the classifier posterior over the items actually selected across all ten sessions.

Two features of the site-level data matter for the pooled conclusions. First, the direction of every condition contrast is the same at all four universities: no site reverses the ordering, and the adaptive–fixed gap ranges narrowly around the pooled estimate. Second, the two national pairs do not differ systematically in their T1 baselines, which is what the within-site randomisation was intended to secure. The site-level tables are therefore consistent with treating site as a stratification variable rather than a factor, while making the evidence base for that decision inspectable.

4.6 RQ3: learner accounts of the feedback loop

Three themes were constructed from the interviews.

- 1) **"It heard something I could not hear."** Seven of the nine learners in the AI conditions described the contour display as making a previously invisible dimension visible. The comparison was frequently drawn not with a teacher but with their own unaided listening: *"When I just listen, I know something is off and that is all I know. Here it tells me the beginning is too high. I would never have got that."* The information that learners reported as most useful was not the tone label but the localisation of the error on the contour — a finding consistent with the cue-weighting account in Section 2.1.
- 2) **"It kept moving to the ones I get wrong."** All five adaptive-condition interviewees spontaneously described the system as tracking them, and two described this as initially uncomfortable: *"It felt like it was hunting me."* The discomfort resolved when learners understood the logic; the two who reported not resolving it both belonged to the lower half of the T2 distribution within their condition. Perceived adaptivity, then, was not automatically motivational; it required the learner to model the system.
- 3) **"I stopped believing it."** The most consequential theme concerned trust. Eight of the nine AI-condition interviewees described at least one episode in which they disbelieved the feedback, and in six cases the trigger was a specific misclassification: the system labelled a Tone 3 token produced in a carrier phrase as Tone 2. Because the citation-form dipping contour is rare in connected speech, learners who compared the system's judgement against a remembered citation form concluded that the system was wrong when it was in fact right about the connected-speech token — or that it was right about a token they had produced differently from what they intended. Either way, the effect was the same: learners discounted subsequent feedback, and three reported deliberately ignoring the contour display for the remainder of that session. Trust was conditional on the system's errors being *legible* to the learner, not on the system being accurate. Figure 5 maps the three themes onto the two constructs they converge on.

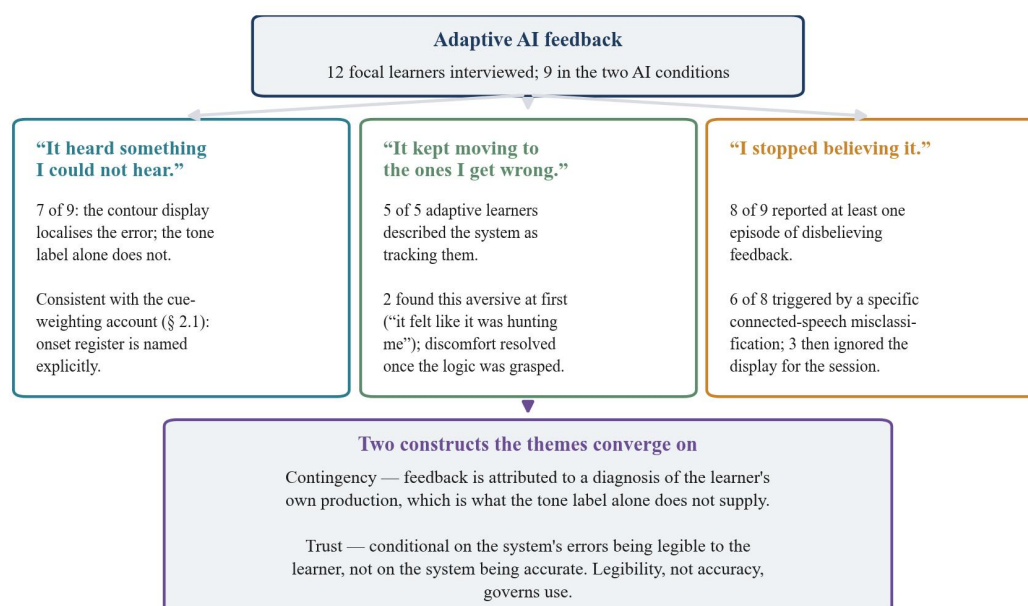


Figure 5. Thematic map of the learner accounts (nine AI-condition interviewees), with the two constructs the themes converge on.

5. DISCUSSION

5.1 Contingency, not the algorithm

The pattern is difficult to explain by appeal to the presence of AI. All three conditions were exposed to the same talkers for the same duration; the two AI conditions received identical feedback components generated by the same model under the same instruction. The only difference was how items were chosen, and yet the adaptive condition outperformed the fixed condition by 10.5 percentage points at both post-test occasions, with a comparable gap in discrimination sensitivity and an equally large advantage on talkers the learners had never

heard. This is the signature of *contingency*, and it aligns with the tutoring literature's central finding that step-level responsiveness, not model sophistication, carries most of the benefit [10, 12]. It also implies that studies contrasting "AI" with "no AI" are uninformative about design: in the present data, a fixed AI condition and an adaptive AI condition differ from each other by about as much as the fixed condition differs from no AI at all.

An important corollary concerns how quickly the adaptive advantage appeared. The selector's moving average over the item bank was crude — a small set of multipliers on error direction and register — and no learner model was fitted. The result is consistent with Sadakata and McQueen's [23] observation that the returns to training design depend on individual aptitude: adaptivity does not need to be optimal to be useful, because its function is to keep the learner inside a region where errors are informative rather than to diagnose the learner precisely.

5.2 Why this contrast in particular

The advantage was of the same size for identification and for discrimination sensitivity, and it did not shrink when the talkers changed. Both facts are informative. Identification is the task in which the learner must commit to a category, so it is the task in which the cue the feedback selects — onset register — is most directly usable; discrimination can be performed on the basis of raw acoustic difference and therefore has less room to benefit from a cue-shifting intervention, so the absence of a large task asymmetry suggests that the intervention changed the perceptual weighting rather than only the response strategy. The transfer result is the more consequential one. The fixed condition's own trained-to-novel drop was steeper than the adaptive condition's, which is why the between-group advantage survived the change of talker essentially undiminished. We read this as evidence that the adaptive condition delivered more tokens from the register in which the learner actually erred — connected-speech Tone 3 realised as a low tone — so that what the learner practised was closer to the input they meet outside the laboratory than a maximally varied but register-blind sequence would be. That is a testable refinement of the high-variability principle: variability of talker is not the only dimension on which training input must vary, and the register distribution may matter as much as the talker distribution.

5.3 The non-target-language environment

Learners in these settings have little ambient input and few opportunities to experience communicative failure that would signal a tone error. The comparison condition's flat trajectory across ten sessions of identical exposure is consistent with the view that mere exposure is insufficient when the perceptual cue is not made salient, and the fixed condition's reliable but smaller gain shows that salience without contingency is only part of the answer. In that sense, the relevant property of the intervention is not its adaptivity in the abstract but the fact that it supplies, in a low-input environment, the kind of contingent information that an immersion setting supplies for free — a point that carries particular weight where, as in the four sites studied here, contact hours are short and ambient Chinese input is close to absent. We would expect the *incremental* benefit of adaptivity to be smaller where communicative pressure already supplies this information, which is a prediction worth testing and not an assumption to be generalised from the present data.

5.4 Trust as a design constraint, not a user attitude

The interview findings recast the technical question of classifier accuracy as a design question about legibility. Learners did not need a perfect classifier; they needed to be able to tell when it was likely to be wrong. Because the citation-form dipping contour is rare in connected speech, the tokens on which the system's behaviour was most likely to be misread by learners were exactly the tokens that carry the most pedagogical value — the connected-speech realisations of Tone 3. Two design consequences follow. First, feedback should display uncertainty rather than only a categorical label, so that a confident decision and a borderline one are distinguishable to the learner. Second, the pedagogical framing of the target should be introduced before the practice, so that learners do not bring a citation-form expectation to a connected-speech task and conclude that the system is broken. Both are pedagogical decisions about the presentation of Chinese, which is consistent with the position that in this domain the object of study is the learner and the language, and the algorithm is the means.

6. LIMITATIONS AND FUTURE RESEARCH

The study has five principal limitations. The first is the sample: it is small and drawn from four institutions in two countries, and it was deliberately restricted — in both countries — to learners whose dominant L1 was Urdu or English. That restriction is what made the contrast interpretable, but it also means the study says nothing about the many learners in both settings whose L1 is a tone language and for whom the Tone 2–Tone 3 problem has a different structure. The second is the follow-up interval of four weeks, which cannot distinguish durable learning from a slowly decaying trace; longer intervals are needed. The third is that the study measured perception, with production only as an underpowered probe; the relationship between adaptive perceptual training and production is the practically most consequential open question, and it cannot be answered here. The fourth is that the study held the feedback content constant across conditions by design, which isolates contingency but also means that the interaction between contingency and feedback *type* — label-only versus metalinguistic explanation — remains unexamined. A factorial design crossing the two would be the natural next step, as would a comparison of adaptive item selection against adaptive *feedback tailoring* with the item sequence held fixed, which would separate the two senses in which an AI system can be described as responsive. The fifth is that site entered the analysis as a stratification variable; a study designed and powered for cross-country comparison could ask whether the contingency advantage itself varies with institutional context, which the site-level tables in Section 4.5 give no more than a first descriptive look at.

A further question concerns teachers. The system in this study operated without a teacher in the loop. Whether the same gains would be observed, and whether teachers would accept the system's classification decisions, are empirical questions that belong to the pedagogical strand of this journal's scope rather than to the perception literature.

7. CONCLUSION

The study asked whether adaptive AI feedback changes how beginners perceive the Mandarin Tone 2–Tone 3 contrast in a non-target-language environment, and whether any advantage is attributable to the adaptivity of the feedback loop rather than to the presence of an AI tutor. Adaptivity produced a substantial advantage over an equally intensive fixed-sequence condition, that advantage was undiminished on talkers the learners had never heard, and it was retained four weeks later; equally intensive exposure without feedback produced essentially no change. Learners' accounts located the benefit in the cue-specific localisation of the error rather than in the tone label, and identified the legibility of the system's errors as the condition for trusting it. The practical implication for Chinese language teaching outside Chinese-speaking environments is that the design property worth investing in is the responsiveness of the loop, and the design property most likely to be neglected is the display of uncertainty on the tokens where the target language itself is variable.

DECLARATIONS

Ethics approval. Approved by the research ethics committees of all four participating institutions, with NUML Islamabad and Nnamdi Azikiwe University acting as lead sites for the two national pairs. All participants gave written informed consent in English or Urdu; Nigerian-site consent materials were additionally glossed in Igbo and Yoruba on request.

Data availability. The complete participant-level dataset and the analysis scripts can be obtained from the corresponding author on reasonable request, and every statistic in Sections 4.1–4.5 is reproducible from them. Audio recordings are not redistributable; see Section 3.7.

Generative AI disclosure. A large language model generated the feedback text delivered to learners in the two AI conditions, under the instruction quoted in Section 3.2. Generative AI was not used to select items, score responses, analyse data, or write this manuscript.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests. The authors declare no competing interests. The tone-classification model used in this study is not a commercial product and no author holds a financial interest in it.

Author contributions. Conceptualization, S.M.A. and A.B.; methodology, S.M.A. and A.B.; software, S.M.A.; validation, S.M.A. and A.B.; formal analysis, S.M.A.; investigation, S.M.A. and A.B.; resources, A.B.; data curation, S.M.A.; writing—original draft preparation, S.M.A.; writing—review and editing, S.M.A. and A.B.; visualization, S.M.A.; supervision, A.B.; project administration, S.M.A. All authors have read and agreed to the published version of the manuscript.

Rights and permissions. The two authors hold the rights to the article.

REFERENCES

- [1] Hao, Y.-C. Second Language Acquisition of Mandarin Chinese Tones by Tonal and Non-Tonal Language Speakers. *J. Phon.* **2012**, *40*, 269–279. <https://doi.org/10.1016/j.wocn.2011.11.001>
- [2] So, C. K.; Best, C. T. Cross-Language Perception of Non-Native Tonal Contrasts: Effects of Native Phonological and Phonetic Influences. *Lang. Speech* **2010**, *53*, 273–293. <https://doi.org/10.1177/0023830909357156>
- [3] Logan, J. S.; Lively, S. E.; Pisoni, D. B. Training Japanese Listeners to Identify English /r/ and /l/: A First Report. *J. Acoust. Soc. Am.* **1991**, *89*, 874–886. <https://doi.org/10.1121/1.1894649>
- [4] Wang, Y.; Spence, M. M.; Jongman, A.; Sereno, J. A. Training American Listeners to Perceive Mandarin Tones. *J. Acoust. Soc. Am.* **1999**, *106*, 3649–3658. <https://doi.org/10.1121/1.428217>
- [5] Wang, Y.; Jongman, A.; Sereno, J. A. Acoustic and Perceptual Evaluation of Mandarin Tone Productions before and after Perceptual Training. *J. Acoust. Soc. Am.* **2003**, *113*, 1033–1043. <https://doi.org/10.1121/1.1531176>
- [6] Barrot, J. S. Using ChatGPT for Second Language Writing: Pitfalls and Potentials. *Assess. Writ.* **2023**, *57*, 100745. <https://doi.org/10.1016/j.asw.2023.100745>
- [7] Godwin-Jones, R. Emerging Spaces for Language Learning: AI Bots, Ambient Intelligence, and the Metaverse. *Lang. Learn. Technol.* **2023**, *27*, 6–27. <https://doi.org/10.64152/10125/73501>
- [8] Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- [9] Kohnke, L.; Moorhouse, B. L.; Zou, D. ChatGPT for Language Teaching and Learning. *RELC J.* **2023**, *54*, 537–550. <https://doi.org/10.1177/00336882231162868>
- [10] Kulik, J. A.; Fletcher, J. D. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Rev. Educ. Res.* **2016**, *86*, 42–78. <https://doi.org/10.3102/0034654315581420>
- [11] Ma, W.; Adesope, O. O.; Nesbit, J. C.; Liu, Q. Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis. *J. Educ. Psychol.* **2014**, *106*, 901–918. <https://doi.org/10.1037/a0037123>
- [12] VanLehn, K. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educ. Psychol.* **2011**, *46*, 197–221. <https://doi.org/10.1080/00461520.2011.611369>
- [13] Chandrasekaran, B.; Sampath, P. D.; Wong, P. C. M. Individual Variability in Cue-Weighting and Lexical Tone Learning. *J. Acoust. Soc. Am.* **2010**, *128*, 456–465. <https://doi.org/10.1121/1.3445785>
- [14] Wong, P. C. M.; Perrachione, T. K. Learning Pitch Patterns in Lexical Identification by Native English-Speaking Adults. *Appl. Psycholinguist.* **2007**, *28*, 565–585. <https://doi.org/10.1017/S0142716407070312>
- [15] Best, C. T.; Tyler, M. D. Nonnative and Second-Language Speech Perception: Commonalities and Complementarities. In *Language Experience in Second Language Speech Learning*; Bohn, O.-S.; Munro, M. J., Eds.; John Benjamins: Amsterdam, 2007; pp 13–34. <https://doi.org/10.1075/llt.17.07bes>
- [16] Flege, J. E.; Bohn, O.-S. The Revised Speech Learning Model (SLM-r). In *Second Language Speech Learning: Theoretical and Empirical Progress*; Wayland, R., Ed.; Cambridge University Press: Cambridge, 2021; pp 3–83. <https://doi.org/10.1017/9781108886901.002>
- [17] Wang, Y.-J.; Li, M.-J. The Effects of Tone Pattern and Register in Perceptions of Tone 2 and Tone 3 in Mandarin. *Acta Psychol. Sin.* **2010**, *42*, 899–908 (in Chinese). <https://doi.org/10.3724/SP.J.1041.2010.00899>
- [18] Wong, P. C. M.; Perrachione, T. K.; Parrish, T. B. Neural Characteristics of Successful and Less Successful Speech and Word Learning in Adults. *Hum. Brain Mapp.* **2007**, *28*, 995–1006. <https://doi.org/10.1002/hbm.20330>
- [19] Francis, A. L.; Ciocca, V.; Ma, L.; Fenn, K. Perceptual Learning of Cantonese Lexical Tones by Tone and Non-Tone Language Speakers. *J. Phon.* **2008**, *36*, 268–294. <https://doi.org/10.1016/j.wocn.2007.06.005>
- [20] Lively, S. E.; Logan, J. S.; Pisoni, D. B. Training Japanese Listeners to Identify English /r/ and /l/. II: The Role of Phonetic Environment and Talker Variability in Learning New Perceptual Categories. *J. Acoust. Soc. Am.* **1993**, *94*, 1242–1255. <https://doi.org/10.1121/1.408177>
- [21] Sereno, J. A.; Wang, Y. Behavioral and Cortical Effects of Learning a Second Language. In *Language Experience in Second Language Speech Learning*; Bohn, O.-S.; Munro, M. J., Eds.; John Benjamins: Amsterdam, 2007; pp 239–258. <https://doi.org/10.1075/llt.17.22ser>
- [22] Perrachione, T. K.; Lee, J.; Ha, L. Y. Y.; Wong, P. C. M. Learning a Novel Phonological Contrast Depends on Interactions between Individual Differences and Training Paradigm Design. *J. Acoust. Soc. Am.* **2011**, *130*, 461–472. <https://doi.org/10.1121/1.3593366>
- [23] Sadakata, M.; McQueen, J. M. Individual Aptitude in Mandarin Lexical Tone Perception Predicts Effectiveness of High-Variability Training. *Front. Psychol.* **2014**, *5*, 1318. <https://doi.org/10.3389/fpsyg.2014.01318>
- [24] Ingvalson, E. M.; Barr, A. M.; Wong, P. C. M. Poorer Phonetic Perceivers Show Greater Benefit in Phonetic-Phonological Speech Learning. *J. Speech Lang. Hear. Res.* **2013**, *56*, 1045–1050. [https://doi.org/10.1044/1092-4388\(2012\)12-0024](https://doi.org/10.1044/1092-4388(2012)12-0024)
- [25] Hattie, J.; Timperley, H. The Power of Feedback. *Rev. Educ. Res.* **2007**, *77*, 81–112. <https://doi.org/10.3102/003465430298487>
- [26] Lyster, R.; Ranta, L. Corrective Feedback and Learner Uptake: Negotiation of Form in Communicative Classrooms. *Stud. Second Lang. Acquis.* **1997**, *19*, 37–66. <https://doi.org/10.1017/S0272263197001034>
- [27] Saito, K.; Lyster, R. Effects of Form-Focused Instruction and Corrective Feedback on L2 Pronunciation Development of /ɹ/ by Japanese Learners of English. *Lang. Learn.* **2012**, *62*, 595–633. <https://doi.org/10.1111/j.1467-9922.2011.00639.x>
- [28] Saito, K.; Plonsky, L. Effects of Second Language Pronunciation Teaching Revisited: A Proposed Measurement Framework and Meta-Analysis. *Lang. Learn.* **2019**, *69*, 652–708. <https://doi.org/10.1111/lang.12345>
- [29] Ellis, R. A Typology of Written Corrective Feedback Types. *ELT J.* **2009**, *63*, 97–107. <https://doi.org/10.1093/elt/ccn023>

- [30] Norris, J. M.; Ortega, L. Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-Analysis. *Lang. Learn.* **2000**, *50*, 417–528. <https://doi.org/10.1111/0023-8333.00136>
- [31] Sweller, J. Cognitive Load during Problem Solving: Effects on Learning. *Cogn. Sci.* **1988**, *12*, 257–285. https://doi.org/10.1207/s15516709cog1202_4
- [32] DeKeyser, R., Ed. *Practice in a Second Language: Perspectives from Applied Linguistics and Cognitive Psychology*; Cambridge University Press: Cambridge, 2007. <https://doi.org/10.1017/CBO9780511667275>
- [33] Kuklick, L. Effects of Learner Choice over Automated, Immediate Feedback. *Learn. Instr.* **2025**, *96*, 102065. <https://doi.org/10.1016/j.learninstruc.2024.102065>
- [34] Chapelle, C. A. The Relationship between Second Language Acquisition Theory and Computer-Assisted Language Learning. *Mod. Lang. J.* **2009**, *93*, 741–753. <https://doi.org/10.1111/j.1540-4781.2009.00970.x>
- [35] Li, Z.; Hegelheimer, V. Mobile-Assisted Grammar Exercises: Effects on Self-Editing in L2 Writing. *Lang. Learn. Technol.* **2013**, *17*, 135–156. <https://doi.org/10.64152/10125/44343>
- [36] Sun, Y.; Kyaw, W. T.; Zhang, J.; Sagisaka, Y. Analysis of L2 Learners' Progress of Distinguishing Mandarin Tone 2 and Tone 3. In *Interspeech 2018*; ISCA, 2018; pp 2549–2553. <https://doi.org/10.21437/Interspeech.2018-1983>
- [37] Wu, Y. The Assessment of Automated Rating of L2 Mandarin Prosody in Lexical Tone Recognition and Pauses. In *Speech Prosody 2024*; ISCA, 2024; pp 250–254. <https://doi.org/10.21437/SpeechProsody.2024-51>
- [38] Chen, M. Computer-Aided Feedback on the Pronunciation of Mandarin Chinese Tones: Using Praat to Promote Multimedia Foreign Language Learning. *Comput. Assist. Lang. Learn.* **2024**, *37*, 363–388. <https://doi.org/10.1080/09588221.2022.2037652>
- [39] Cao, M.; Pavlik, P. I.; Bidelman, G. M. Enhancing Lexical Tone Learning for Second Language Speakers: Effects of Acoustic Properties in Mandarin Tone Perception. *Front. Psychol.* **2024**, *15*, 1403816. <https://doi.org/10.3389/fpsyg.2024.1403816>
- [40] Godwin-Jones, R. Distributed Agency in Second Language Learning and Teaching through Generative AI. *Lang. Learn. Technol.* **2024**, *28*, 5–31. <https://doi.org/10.64152/10125/73570>
- [41] Law, L. Application of Generative Artificial Intelligence (GenAI) in Language Teaching and Learning: A Scoping Literature Review. *Comput. Educ. Open* **2024**, *6*, 100174. <https://doi.org/10.1016/j.caeo.2024.100174>
- [42] Wang, Y.; Zhang, T.; Yao, L.; Seedhouse, P. A Scoping Review of Empirical Studies on Generative Artificial Intelligence in Language Education. *Innov. Lang. Learn. Teach.* **2025**. <https://doi.org/10.1080/17501229.2025.2509759>
- [43] Zawacki-Richter, O.; Marin, V. I.; Bond, M.; Gouverneur, F. Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. <https://doi.org/10.1186/s41239-019-0171-0>
- [44] Mizumoto, A.; Eguchi, M. Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. *Res. Methods Appl. Linguist.* **2023**, *2*, 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- [45] Hanif, B. Current Scenario and Perspective of Teaching Chinese at Confucius Institutes in Pakistan. *Pak. J. Humanit. Soc. Sci.* **2023**, *11*, 2380–2391. <https://doi.org/10.52131/pjhss.2023.1102.0530>
- [46] Olayoku, P. A. Mandarin Education for Economic Empowerment: The Confucius Institute in Lagos, Nigeria. In *Mobility between Africa, Asia and Latin America: Economic Networks and Cultural Interactions*; Röschenthaler, U.; Jedlowski, A., Eds.; Zed Books: London, 2017; pp 236–258. <https://doi.org/10.5040/9781350221406.ch-012>
- [47] Plonsky, L.; Oswald, F. L. How Big Is “Big”? Interpreting Effect Sizes in L2 Research. *Lang. Learn.* **2014**, *64*, 878–912. <https://doi.org/10.1111/lang.12079>