Zhang and MacWhinney
Language Testing in Asia (2023) 13:35
https://doi.org/10.1186/s40468-023-00232-6

Language Testing in Asia

RESEARCH Open Access

Using diagnostic feedback to enhance the development of phonetic knowledge of an L2: a CALL design based on the unified competition model and the implementation with the Pinyin Tutor

Yanhui Zhang[1*] and Brian MacWhinney[2]

*Correspondence: Yanhui.Zhang@nottingham.edu.cn

[1] School of Education and English, University of Nottingham Ningbo China, Ningbo, China
[2] Department of Psychology, Carnegie Mellon University, Pittsburgh, USA

Abstract

Second language acquisition (SLA) is complex and multidimensional. Using the framework of the unified competition model (UCM), the current study explores how robust learning and testing of Chinese Pinyin are fostered by optimal integration of different kinds of feedback in an intelligent computer-assisted language learning (CALL) environment offered by the Pinyin Tutor at TalkBank. The findings demonstrated that the repeated feedback-embedded training with the Pinyin Tutor significantly boosted the learners' proficiency in all aspects of Pinyin knowledge for second language (L2) learners of Chinese whose first language (L1) backgrounds were varied and whose initial proficiencies in Chinese were elementary. Furthermore, there was a strong increase in Pinyin knowledge, as evidenced in the delayed posttest administered 3 months after finishing the training sessions. The results further showed that diagnostic feedback led to greater improvement than basic feedback. The significance of the results is attributed to the design of the Pinyin Tutor, which implements principles from psycholinguistic theory as well as corpus data on the speech production by L2 learners. The study sheds fresh light on improving the Pinyin Tutor, and CALL in general, by incorporating up-to-date findings in educational psychology.

Keywords: Diagnostic feedback, CALL, Contrastive analysis, Pinyin Tutor, Unified competition model, Language assessment, Phonetic knowledge development

Introduction

As hypothesized by the unified competition model (UCM), second language acquisition (SLA) is a complex and multidimensional process, involving the dynamic interactions of a large number of forces and constraints with different magnitudes operating across various timescales (MacWhinney, 2012, 2018, 2021). According to the UCM, the main objective of language education, assessment, and curriculum design should be to foster various protective factors while mitigating various risk factors, with detailed individual

differences being taken into account (MacWhinney, 2015, 2018; Pawlak, 2019, 2022). The UCM emphasizes parallels between first language (L1) acquisition and second language (L2) acquisition. However, what is different for these two types of learners is the dramatically different learning environments they are facing and, correspondingly, the vastly varied substantialities and contributions of all the protective or detrimental factors in shaping overall learning performance (MacWhinney, 2018).

There is no doubt that acquistion of a native-like accent is more difficult for L2 learners who begin to learn L2 after puberty (Hartshorne et al., 2018; Guion et al., 2000; Flege, 1995; Patkowski, 1990; Johnson & Newport, 1989; Hernandez et al., 2005). This is perhaps even more true for the acquisition of Chinese pronunciation of both segments and tones. Studies show that the attainment of native proficiency in Chinese pronunciation, including both tones and segments, is extremely difficult for L2 learners of Chinese who begin learning after childhood (Hu, 2010; Qi et al., 2015; Stickler & Shi, 2013). While there have been a variety of learning models (e.g., contrastive analysis by Lado, 1957; markedness differential hypothesis, Eckman, 1981, 1991) attempting to interpret this age-related phenomenon, the UCM emphasizes the roles of (1) the entrenchment of auditory perception and motor forms, (2) overload of neurocognitive capacities of adult learners by irrelevant or competing tasks, and (3) the lack of the instructional and social support in comparison with the young learners in learning an L1. The problems faced by L2 learners are gradient rather than absolute, as there have been empirical studies indicating some adult L2 learners can achieve native-like pronunciation (Hartshorne et al., 2018; Huang, 2015).

Input and training are indispensable for achieving native-like speaking in L2, as this is fundamentally an entropy-reversing process where various risk factors must be constantly overcome at every stage of learning. Feedback has been argued as one of the most critical learning strategies across all learning modes, regardless of classroom instruction or remote learning, and for the acquisition of all categories of linguistic abilities, including reading, writing, and speaking (Gebril, 2021; Li, 2020; Bitchener, 2008; Kang & Han, 2015; Lyster, 2015). Zhang (2020) demonstrated that, in the long run, reflective feedback is more effective than corrective feedback and rule-based feedback in helping the L2 learning of Chinese classifiers. Li (2010, 2015), for example, provided a meta-analysis on feedback research, showing that, with both oral and written samples under investigation, explicit feedback worked better than implicit feedback over a short time. A positive effect on SLA learning is critically dependent on feedback quality, which may include considerations such as the feedback's contents and design, types, and timing (Li & Roshan, 2019; Kang & Han, 2015; Zhang & Li, 2016; Fu & Li, 2022; Li, 2020).

A robust CALL platform adopted in the current study is the Pinyin Tutor developed and maintained under TalkBank at Carnegie Mellon University (www.TalkBank.org). In correcting the shortcomings of many existing CALL systems, the design of the Pinyin Tutor has been based on the authentic learning corpus of thousands of L2 learners' learning profiles and datasets. Learners' error patterns in the tutor were rigorously analyzed and categorized based on the theoretical phonological contrast analysis between Chinese and other major languages such as English, Korean, and Cantonese (Kowalski et al., 2014). Data on error patterns were updated as new data streamed in from all the registered learners across more than one-hundred colleges and universities across the

USA and other countries worldwide. This corpus-driven and learner-centered design made it possible to investigate how learning can be optimized in the CALL environment. Recent research (Zhang & MacWhinney, 2023) provided strong evidence that the Pinyin Tutor, by adopting training stimuli with differentiated levels of novelty, can significantly enhance the acquisition of Chinese phonetic knowledge among L2 learners of Chinese with varied proficiencies and backgrounds.

When benchmarking the robustness of the Pinyin Tutor, it is crucial to examine the effects of feedback. Feedback helps to boost the cue strength and resonant connections of the orthographic and auditory forms of a word while suppressing competition from incorrect alternatives (MacWhinney, 1991). Phonology-orthography mappings provide a significant source of resonance in L2 learning for alphabetic languages. However, language learners with Roman scripts find it difficult to utilize the same source when learning Chinese through Hanzi characters (MacWhinney, 2008). The Pinyin Tutor introduced in the current study helps to overcome such challenges by relying on Pinyin, rather than Hanzi. The training offered by the tutor is expected to show an advantage in learning the mapping of oral Chinese to its graphic representation. We expect that feedback should work together with resonant practice to enhance and consolidate the learning of Chinese as an L2.

Theoretical framework and literature

Researchers have proposed several major theories to understand the development of L2 phonological categories and processes. One prominent theory for L2 speech development and learning is the native language magnet (NLM) model (Kuhl, 1991, 1998, 2000). The NLM model assumes a fixed developmental sequence of speech perception from infant to adult, i.e., from language-universal phase to language-specific. Infants map acoustic sounds from the input of the environment and gradually form the specific speech processing system, subserving a filter for the L1 perception and production. Accordingly, the phonetic category of the infant's first language functions as the prototype of the infant's neurophonological system, magnetizing the nearby constituents to the formed category. Once this language-specific mental network is formed, the acquisition of L2 phonemes becomes difficult as the L2 sound does not accurately correspond to any categorized mappings in such an existing network to give rise to the accurate reproduction of the L2 input. The main pitfall of the NLM is its overestimation of the L1 entrenchment and lack of attention to the influence of instructional strategies and language input on the attainable level of SLA (Flege, 2018; Pawlak, 2019; MacWhinney, 2018).

Two other major approaches to the learning of L2 phonology are the perceptual assimilation model (PAM) (Best, 1994, 1995; Best & Tyler, 2007) and the speech learning model (SLM) (Flege, 1995). Both models concern how similarity in sounds between two languages may affect L2 acquisition, especially at the early stage of L2 learning. The PAM assumes that adult listeners with developed phonological categories perceive L2 phonemes by the similarities to their L1 phonemes (Best, 1994). The L2 listener would find it difficult to distinguish the differences between L1 and L2 phonemes if they have very similar phonological properties. Therefore, what is most likely to occur during the learning is that the listener will assimilate the L2 phoneme to the L1 category that is

perceived to be most similar. The SLM emphasizes that whether an L2 phonetic category can be successfully formed depends on the phonetic distance of the L1 and L2 sounds. The model also observes that speech perception ability changes over the ages of the learner in that learner of older ages typically perceives and produces L2 articulations less accurately than the younger.

The revised version of the SLM (Flege & Bohn, 2021) extends the SLM in ways that are highly compatible with the UCM. Both theories emphasize the role of amount of L2 input, rather than simply the numbers of years in the L2 community, and both theories emphasize the critical roles of individual difference factors in motivation and L1 ability. Like the UCM, the SLM attributes contrastive effects to the interaction and competition between L1 and L2 systems. In addition to these large areas of agreement between SLM and UCM, the UCM focuses on the role of the psychological processes of entrenchment, resonance, chunking, and overanalysis and the ways in which risk factors and protective factors vary across linguistic levels, including but not limitedd to phonology. For the purposes of the current study, the most important dimensions of the UCM are those that focus on the role of feedback, resonance, and entrenchment (MacWhinney, 2015, 2018, 2021; Zhang & MacWhinney, 2023). The UCM holds that, particularly for adult learners, explicit and reflective feedback can be particulary effective if they make use of simple cues to form-function mappings. Such cues can help learners hold patterns in memory through resonant connections between orthography, lexicon, and phonology in ways that then promote initial entrenchment.

The classical competition model (MacWhinney, 1987, 1991) was centered around the notion of the competition between various options or cues. The selection between the competing cues is largely determined by the relative cue strength, which, in turn, is a function of cue validity. Cue validity, encapsulating cue availability, cue reliability, and conflict reliability, is operationally defined through the corpus counts of the cue. It is hypothesized that availability is the primary determinant at the initial stage of the acquisitional process. The learner initially decides which cues should be associated with which meanings by using cue availability: a highly available cue will be more likely assigned with high cue strength early in language acquisition. Therefore, L2 learning is expected to be more robust if an L2 learner's attention can be effectively guided by valid cues from the onset.

As with the extended version of the competition model, the UCM includes the analysis of the detailed dynamic and evolving pattern of the learning process instead of focusing only on the end state of language acquisition at specific times. Thus, timescales play essential roles in the UCM when accounting for the effects of protective or supportive factors in SLA study and the different instructional strategies needed to align with these timescale variations (MacWhinney, 2015, 2018, 2021). Such extension in the UCM also allows for the accommodations of highly complex, nonlinear, and stochastic language phenomena, such as self-organization and bootstrapping, into the framework of the model. At the core of the UCM are dynamic and interconnected competing processings shaped by constraints that vary across timescales and linguistic levels (Caldwell-Harris & MacWhinney, 2023).

A fundamental implication of the UCM is that competition can be shaped by learning and instructional strategies, such as feedback. Feedback may foster positive

interlanguage transfer when L2 learning is conceptualized as the process of establishing an additional set of form-function relationships in a target language. For example, forms unmarked in L1 are usually transferred more strongly than marked forms (Eckman, 1991; Major, 2001). A typical example is the difficulty of native Japanese speakers to distinguish between English /r/ and /l/ in perception. MacWhinney (2008) noted that the transfer of L1 sounds to L2 is the cause of strong L1 accents in L2, and that such transfer is counter-productive in the long run because it "embeds L1 phonology into the emergent L2 lexicon" and results in "long-term difficulties in correcting entrenched erroneous phonological transfer." Therefore, to promote the acquisition of marked L2 phonemes, it is essential to optimize the input quality and frequency and provide cue-focused feedback.

In online settings, learners need to rely even more on high-quality feedback. Although CALL provides unprecedented convenience in terms of learners' autonomy, self-efficacy, and self-regulated incidental learning, it has inherent shortcomings, including the absence of real-time interactions with teachers and the lack of social and physical support that a classroom environment can offer (Tsou et al., 2006; MacWhinney, 2015; Lyster, 2015). From the UCM perspective, a productive CALL environment must promptly pinpoint errors and items to improve throughout various learning stages to strengthen the valid cues, mappings, and resonances in the most effective manner. Also, L2 instruction should calibrate the level and depth of feedback to optimize the resonance and internalization. This can be done using tailored task ladders for drill and self-contained hints and explanations to foster the learner's metacognition and motivate engagement (MacWhinney, 2015). In other words, CALL and feedback do not automatically bring about effective learning. Whether and to what extent feedback in CALL may lead to improved learning depends on whether the feedback has been optimally matched with the learners' individual needs. Given that the empirical studies towards such quest are rather scarce to date, the current study aims to fill the gap by investigating which feedback condition, basic or diagnostic, is more conducive to SLA in CALL.

Specifically, the objective of the current study is to examine how different types of feedback may differently affect the development of Chinese phonological skills among L2 learners of Chinese in an innovative CALL environment provided by the Pinyin Tutor at TalkBank, an online human–machine interactive L2 learning platform. The current study aims to answer the following research questions.

1. To what extent will the repeated practice through the Pinyin Tutor promote positive L2 learning of Pinyin?
2. To what extent will L2 learners benefit from the diagnostic or basic feedback provided by the Pinyin Tutor?

Implementation with the Pinyin Tutor

Programmed in Java and hosted at the TalkBank (MacWhinney & Fromm, 2022), Carnegie Mellon Unversity, the Pinyin Tutor is a web-based Chinese Pinyin learning and assessment platform suited for both self-regulated training and curriculum-oriented classroom instructions. The Pinyin Tutor trains the learners to spell a word and provide

corresponding feedback at either minimum or diagnostic levels. If the word is spelled incorrectly three times, the tutor will present the correct spelling to the learner and then proceed to the next item. When the learner finishes the dictation task, the tutor provides the accuracy rate and asks if the learner would like to redo the incorrect items: "Your score is: x%. Would you like to review the items you missed?" If the learner selects "yes," the tutor will extract the items spelled incorrectly and let the learner continue practicing these problematic items. Otherwise, if the learner chooses "no," the tutor will quit by displaying "zai4jian4" (goodbye). As learners practice on the Pinyin tutor, their performances are recorded by the system. Authorized researchers and instructors can view the log files, trace the learning progress, and identify specific learning problems. Log file data include training-specific information such as user alias, score, time and length of training, the round of practice, number of attempts in each item, and stimuli. The portfolio is updated as soon as a learner completes a training session. Instructors can monitor the performance of their students both individually and as a group to better understand the difficulties posed in both the teaching and learning of Pinyin for individual students.

The Pinyin Tutor was designed to implement instructional strategies and methods consistent with the UCM framework. Although the experiment reported in the current study focuses on the impact of feedback on the L2 acquisition of Chinese phonological knowledge, the tutor provides a variety of additional features aligned with the UCM, such as the option of using familiar or novelty cues and tailored scheduling of practices. In addition, the design and the underlying algorithm allow easy extensions to learning other languages or other dialects of the same language. Second, the Pinyin Tutor has the capacity to automatically populate and analyze the data generated throughout training, particularly those pertinent to learning performance and progress, providing a thorough, reliable, and rigorous reference for classroom instructors to optimize their teaching plans and instructional methods. The tutor can generate learning behavior reports covering a large-scale corpus at various timescales, which is helpful for educational assessment, modeling, and policy-making purposes at a macro level (Kowalski et al., 2014). Third, the feedback databases embedded in the Pinyin Tutor design were based on rigorous data-mining analysis of the large-scale authentic corpus of L2 speakers with diversified L1 backgrounds after being benchmarked with the theoretical contrastive phonological analysis between Chinese and other major languages. For completeness, all the pronounceable syllables of Mandarin Chinese, close to four thousand in total, were individually calibrated and included in the underlying database of the Pinyin Tutor. Figure 1 is an illustration of the conceptual design of the main interface of the tutor.

A learner may register as an individual user for self-regulated learning. The instructor for classroom instruction can create a group account or a link to each training session. Upon navigation to the login page, students are required to enter their registered credentials to log in to the platform. The system double-checks the correctness of the IDs to avoid misspellings. After logging in, the screen will remind the student to check the audio volume. Descriptions and tutorials are available, informing the learners on how to interact with the tutor and how to enter the Pinyin in the text box. For instance, users are instructed to use the numbers 1, 2, 3, and 4 for the four tones and the number 5 for the neutral tone. Also, spaces between syllables do not affect spelling accuracy. More specific instructions and examples are provided in Zhang and MacWhinney (2023).

Fig. 1 Conceptual design of the main interface of the Pinyin Tutor

The learner can press the "start" button to start a Pinyin dictation practice. An open field for Pinyin entry, four navigating buttons, and one field for feedback are then presented on the web page for a Pinyin dictation task. After listening to a Chinese word, the learner can enter its Pinyin transcription. Then, the learner has the alternative of clicking buttons to check the spelling or moving on to the next item. Specifically, the "Listen to Target" button allows learners to listen to the target pronunciation repeatedly; the "Listen to Your Attempt" button is grayed before the learner types in any answers in the open field. When an error occurs, if the entry is regarded as a pronounceable item, this button would be made available so the learner can listen to the word he/she had spelled. The "Check" button informs the learner whether an item is correct or not. The "Next" button allows learners to proceed to the next Pinyin dictation item.

In the basic feedback condition, the Pinyin Tutor only indicates if a Pinyin spelling is correct or not. In the diagnostic feedback condition, detailed feedback is provided when an error occurs. Examples of the diagnostic feedback are presented in Table 1. The pool of diagnostic feedback messages was constructed based on a thorough error analysis of the historical learners' corpus from contrastive analysis and universal markedness perspective. The analysis examined the corpus for all the pronounceable syllables of Chinese, their common error types, and their difficulties regarding the feature system, segment inventory, and the distribution of errors. For instance, English has a set of affricates that sound like the Chinese retroflex affricates but are prone to be confused by L2 learners of Chinese with L1 in English. Furthermore, English has both "r" and "l" in the initial position of a syllable, posing a challenge for English

Table 1 Example of the diagnostic feedback coded in the Pinyin Tutor

Target	Error	Feedback
le4	ne4	Your attempt at the initial, n, is not correct
	la4	Your attempt at the final, a, is not correct
	le2	Tone 2 is not correct
	e4	You did not type the initial
an1	lan1	The initial you typed is not part of the syllable
	an	This is not a neutral tone
yuan3	yvan2	When "v" is used as an initial, change v to u, and put y in front of u
diu1	diou1	Your attempt at the final, iou, is not correct. When -uen/-uei/-iou is pronounced, write it as -un/-ui/-iu
xia4	xa4	Your attempt at the final, a, is not correct. Pinyin writes this as xia4, not xa4. j/q/x only comes before i or u
	sia4	Your attempt at the initial, s, is not correct. Pinyin writes this as xia4, not sia4. zh/ch/sh/r/z/c/s does not come before i as a glide
qi4hou4	qi2huo3	qi2: Tone 2 is not correct
		Your attempt at the final, huo, is not correct
		huo3: Tone 3 is not correct

background learners to listen and speak the Chinese syllable "r". This is an example of the negative transfer from the UCM point of view since the Chinese "r" and English "r" are not identical, because the "r" in Chinese is pronounced with the tongue curled up. Consequently, not hearing the differences between the Chinese "r" and "l", the English L1 learners may associate the pronunciation of "r" with "l," resulting in a mismatch in the Pinyin practice. All such errors were analyzed and coded into the diagnostic feedback messages in the Pinyin Tutor.

Test method and procedure

A series of dictation tasks were applied to train the ability to transcribe Chinese words in Pinyin, with the degree of feedback being the control condition. To transcribe words or sentences, one must decode the auditory input, map phonemes with their corresponding graphemes, and encode them into text. The cognitive process of dictation is illustrated in Fig. 2. Both problems in perception and trouble with the use of the spelling rules cause

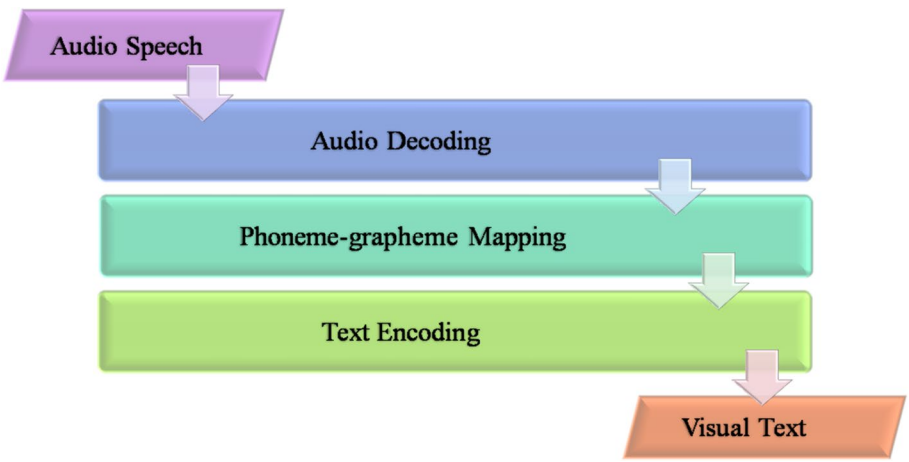

Fig. 2 The cognitive flowchart of the Pinyin dictation task

errors in Pinyin dictation. A proficient L2 learner of Chinese is expected not to be bound by the sound-letter correspondence in his or her native language, and the learner should have a linguistic awareness of phonological structure and the mastery of the phono-orthographical mappings in Chinese Pinyin. Thus, the learner's phonological perception proficiency is predictable by the learner's orthographic transcription accuracy.

Two feedback conditions at different contrasting degrees were designed and embedded in the Pinyin training offered by the Pinyin Tutor, one with only basic feedback on whether a Pinyin transcription was correct or not, and the other with diagnostic feedback regarding how and why the error was made. For example, errors caused by incorrect transcription encoding could be addressed by verbal text feedback on Pinyin spelling rules and descriptions. Detailed verbal feedback was also provided to correct the mismatches due to L1 and L2 phonological problems, Pinyin spelling problems, or both. In both conditions, the tutor allowed the participants to play the correct target sound for comparison with the sound of the orthography they had generated. The target-attempt speech discrimination based on the "minimal pair" principle was adopted in giving feedback. The minimal pair method has been widely used for speech therapy as well as considerable speech perception or production training programs (Bradlow et al., 1997; Gibbon et al., 1997). The contrast applied in the current experiment only represents a minimal pair when the student's production is incorrect in a single dimension. The contrast is non-minimal if the production is wrong in more than one dimension.

To examine the role of feedback experimentally, 83 students enrolled in a regular course in elementary Chinese at an American university participated in the test. The learners received 2 weeks of conventional classroom instruction and learning focusing on Pinyin knowledge before the start of the Pinyin Tutor training sessions in the following week after the Pinyin topic was covered. The participants were randomly allocated to the two training conditions. Depending on whether the total number of letters in the campus ID string was an odd or even number, the computer program assigned the participant to either the diagnostic feedback or minimum feedback training group. The participant stayed in the same treatment group throughout all the sessions so that all participants received the same training stimuli except the types of feedback, which are group specific. Researchers were only allowed to view the reports and analyze the data with anonymized IDs. The number of subjects in the following analysis varied, depending on the nature of the hypothesis. If the score in the posttest was more than 5 points lower than the pretest, data from that participant was dropped. As a result, the test scores of two students were removed from the study by the screening. The participants attended elementary Chinese classes 4 hours per week throughout the semester per credit requirement. None of them reported any hearing disability. Among them, 77 participated in at least one Pinyin training, 55 participated in at least 5 Pinyin sessions of training, and 22 attended both the pre-and posttests and eight training sessions. Data analysis mainly focused on the 55 students who participated in at least five Pinyin training sessions. The distribution of participants' native languages is Korean (32.73%), English (25.45%), Mandarin as a heritage language (16.36%), Cantonese as a heritage language (18.18%), and others (7.27%).

The first phase of the study consisted of a pretest, a posttest, and eight training sessions. The follow-up study in the second phase was a delayed posttest to examine performance

retention. The pretest comprised 40 stimuli, including 17 monosyllabic words, 20 bisyllabic words, and three multisyllabic words. Among them, 16 were unknown words that were not included in the textbook. Participants were given only one opportunity to listen to every word and were asked to transcribe them into Pinyin scripts. The eight training sessions spread from the 5th to the 14th week. Pinyin practice was assigned weekly or biweekly in the form of online homework. The first three training sessions were comprised of monosyllabic words. Stimuli in the remaining five sessions were bisyllabic and multisyllabic words. In the eight training sessions, two-thirds of the stimuli was known words (words from the textbook), and the remaining one-third was novel words (words beyond the textbook). During each training session, participants were given three opportunities to transcribe the Chinese words into Pinyin in the first round. As the first round concluded, participants were given additional opportunities in the subsequent rounds to voluntarily practice the words that they had just misspelled until all the words were correctly spelled. The posttest was administered in the last instructional week. The stimuli and format for the posttest were exactly the same as for the pretest. Together with the posttest, an additional online survey about the participants' language background was completed. The delayed posttest was carried out 13 weeks after the posttest. The stimuli and format were the same as for the pretest and posttest. The timescales of the experiment, spanning 28 weeks throughout the training sessions and tests, are demonstrated in Fig. 3.

A Perl parser partitioned all the training stimuli and the test items into initials, finals, and tones to conduct fine-tuned analysis. Java and Python are the primary programming languages applied to analyze the learning profiles, such as the count of clickings on various buttons, error rate, acquisition rate, and times spent on tasks. When calculating the total times, only continuous activity was considered. When the Pinyin Tutor had been inactive for a period of 3 minutes, this time was not counted into the total duration of practice, allowing for circumstances such as the participant walking away in the middle of Pinyin practice.

Results and analysis

The first research question concerns the overall learning-enhancing effect of the Pinyin Tutor in benefiting the L2 acquisition of Pinyin skills under either of the two feedback modes embedded in the Pinyin Tutor training. The question was quantitatively examined through the score improvement from pretest to posttest and from pretest to delayed posttest as arranged through the training phases. With other conditions held equal and random effects assumed normal, the significance of the Pinyin training effect is statistically equivalent to the score improvement being greater than zero at a reasonable confidence level. For this purpose, paired *t*-tests were run to test the null hypothesis that the

Fig. 3 The timescales of the training sessions and the tests of the experiment

mean scores of the pretest and posttest, at each aspect of Pinyin knowledge, are equal. For these tests, each of the *p*-values is well below 0.001. More specifically, the *p*-values, with sample size $N=42$, for the pretest–posttest score comparison in word, syllable, syllable without tone, initial, final, and tone are, respectively, 9.0095e-09, 8.0226e-13, 201034e-11, 1.2909e-6, 4.0786e-13, and 7.1865e-04. The corresponding effect sizes in terms of Cohen'*d* values are, respectively, 6.9725, 9.9695, 8.8833, 5.4535, 10.2009, and 3.4180. All these effect sizes are well above the benchmark values suggested by Plonsky and Oswald (2014) for an effect size to be significant in linguistic and educational studies. Since the *t*-test is robust to normality, especially for moderate and large samples such as $N>40$ (Rawlings et al., 2006), a simple z-score test can also be applied. Indeed, the *p*-values generated by the z-score procedure tended to be less conservative in general, where, for instance, the *p*-value of z-score test for the improvement in tone is about 0.0003, which is more than 50% smaller than that by the *t*-test. To sum up, there is sufficiently significant evidence, based on the paired z-score test for approximation or *t*-test for statistical rigor, that the Pinyin Tutor benefited the participants' learning in every component of the Pinyin knowledge towards a higher proficiency in Chinese as an L2.

A similar conclusion can be drawn for the comparison between the pretest and the delayed posttest, as all the *p*-values generated from the *t*-test are significant at a 95% confidence level. Specifically, as shown in Table 2 and Fig. 4, the learners' scores in the pretest and delayed posttest, with sample size $N=21$, saw a standardized increase of 21.36, 19.40, 17.10, 8.76, 13.01, and 11.52, respectively, in word, syllable, syllable without tone, initial, final, and tone, where the standardized score was obtained by transforming the raw number of the correct trials of the Pinyin stimuli in the test into the corresponding value in one-hundredth scale assuming equal weights for each stimulus. The corresponding *p*-values for these increments are, respectively, 6.0315e-9, 2.2855e-5, 1.1420e-5, 9.3113e-6, 1.6376e-6, and 1.7841e-2. The corresponding effect sizes in terms of Cohen'*d* values are, respectively, 9.2251, 5.1780, 5.3745, 5.5746, 6.3654, and 2.2526.

The difference between the two mean score increases, from pretest to posttest vs. pretest to delayed posttest, is very small, as shown in Fig. 4 and Table 2. More rigorously, the minimum *p*-value from the formal *t*-tests for comparison of these two means at all six components of Pinyin knowledge is above 0.30, thus confirming the improvements from the pretest to the posttest and from the pretest to the delayed posttest are not statistically different. Such a finding tends to support the long-horizon retention of the Pinyin knowledge learned through the Pinyin Tutor, as the time ellipsed between the posttest

Table 2 Overall learning enhancing effect of the Pinyin Tutor ($N=42$ for pre-post and $N=19$ for pre-delayed post)

	Pre-post			**Pre-delayed post**		
	Mean	**SD**	***p*-value**	**Mean**	**SD**	***p*-value**
Word	17.3474	16.1239	9.01e-09	21.3563	10.6087	6.03e-09
Syllable	18.6217	12.1051	8.02e-13	19.4047	17.1733	2.29e-05
Syllable w/o tone	17.7983	12.9846	2.10e-11	17.1029	14.5828	1.46e-05
Initial	7.5435	8.9644	1.29e-06	8.7566	7.1983	9.31e-06
Final	14.1237	8.9729	4.08e-13	13.0079	9.3646	1.64e-06
Tone	9.1127	17.2783	7.19e-04	11.5241	23.4454	1.78e-02

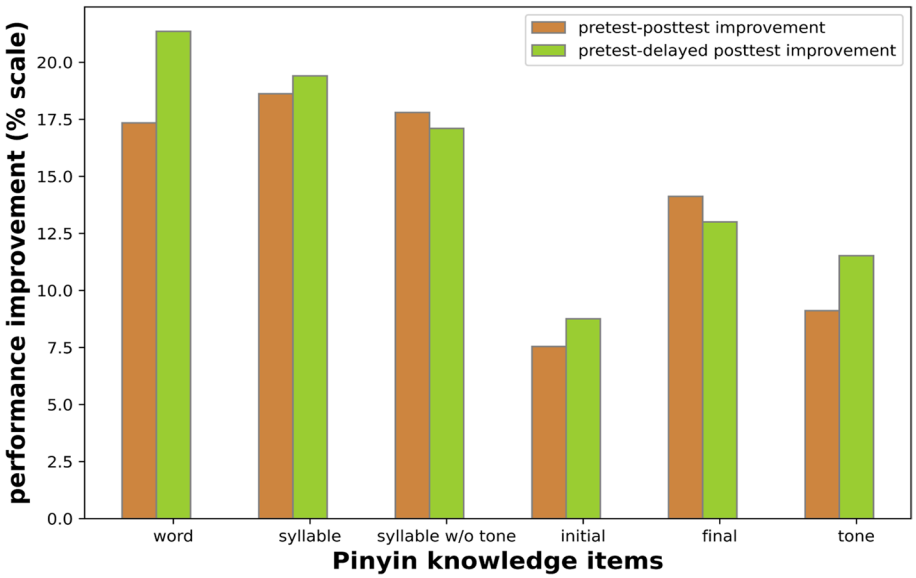

Fig. 4 Pinyin knowledge performance improvements

and delayed posttest was 3 months, a sufficiently long horizon for fundamental phonological skills such as Pinyin listening and spelling to wane if they are not substantialized from neurocognitive theory perspective (see Munro et al., 2012, for instance, for a more detailed discussion on short memory of linguistic knowledge). In summary, the results demonstrated that Pinyin Tutor has significantly enhanced the L2 learners' learning in all aspects of Pinyin knowledge. The retention rate of the acquisition of Pinyin knowledge through the Pinyin Tutor has been proven satisfactory and sustainable as tested through both the posttest and the delayed posttest.

The second research question concerns the relative effectiveness of the two types of feedback, namely, diagnostic vs. basic feedback, in enhancing the L2 learners' acquisition of the various aspects of Pinyin skills. Such effects were tested through the score differences between the pretest and posttest and those between the pretest and the delayed posttest, as taken by the participants at different stages of practice. Table 3 demonstrates the main results for the performance difference between the pretest and posttest, with sample size $N = 23$ for the basic feedback group and $N = 19$ for the diagnostic. Mean improvements for the diagnostic group were systematically

Table 3 Pretest–posttest improvements for the two feedback conditions ($N = 23$ for basic and $N = 19$ for diagnostic)

	Diagnostic		Basic		Significance	
	Mean	SD	Mean	SD	Cohen's *d*	*p*-value
Word	44.3673	19.8887	23.6174	16.3091	3.6446	0.0004
Syllable	39.6408	11.0356	33.3367	11.3375	1.8201	0.0381
Syllable w/o Tone	36.0326	13.0538	33.8581	7.2024	0.6483	0.2602
Initial	15.3227	8.4316	14.1964	6.9511	0.4658	0.3220
Final	26.5857	10.9375	24.1430	7.1932	0.8348	0.2044
Tone	26.9685	19.1940	9.4838	14.8712	3.2446	0.0012

higher than those for the basic group. Hence, diagnostic feedback, on average, generally helped the Pinyin learning through practice on the Pinyin Tutor for all the tested Pinyin knowledge items. There is no evidence of a floor effect, as the pre-training proficiencies of the participants from the two groups did not exhibit statistical difference with p-value > 0.70 for the F-test for comparison of means of their pretest scores. The advantage of diagnostic feedback in learning enhancement is particularly significant at the tone level of Pinyin knowledge since the average score increment between pre-post tests for the diagnostic feedback group of learners is 26.97, as opposed to 9.48 for the basic feedback group, which corresponds to a two-tailed p-value of 0.001 from one-way ANOVA.

Given that the classical comparison of means in ANOVA assumes relatively stricter conditions on the samples, different kinds of p-values were calculated and presented in Fig. 5 to mitigate the potential biasedness. Specifically, the Fisher permutation test and Wilcoxon rank-sum test were chosen for comparison in consideration of their lower sensitivity to normality and sample size. As demonstrated, the three kinds of p-values for testing the significance of the advantage of diagnostic feedback produced consistent results. However, the p-values from the classical ANOVA t-test are more significant than those from the other two. The results showed that the learning-enhancing effect of diagnostic feedback is more decisively higher than basic feedback in the levels of word, syllable, and tone, where, for instance, the p-values for the Fisher permutation test with 1000 iterations, as shown by Fig. 6, are 0.001, 0.054, and 0.057, respectively, all statistically significant at a 90% level of confidence. On the other hand, the advantage of diagnostic feedback over basic feedback in terms of pre-post test score improvement is less decisive in the levels of syllable without tone, initial, and final, where, for instance, the p-values from the ANOVA t-test are, respectively, 0.260, 0.322, and 0.204. Similar results are shown from the other two kinds of

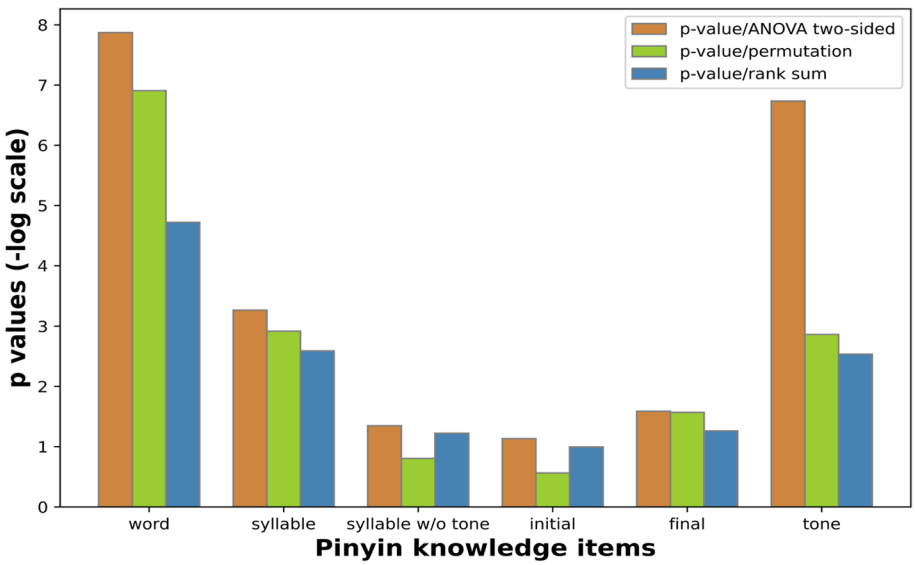

Fig. 5 Significance of pretest–posttest improvements for the two feedback conditions ($N = 23$ for basic and $N = 19$ for diagnostic)

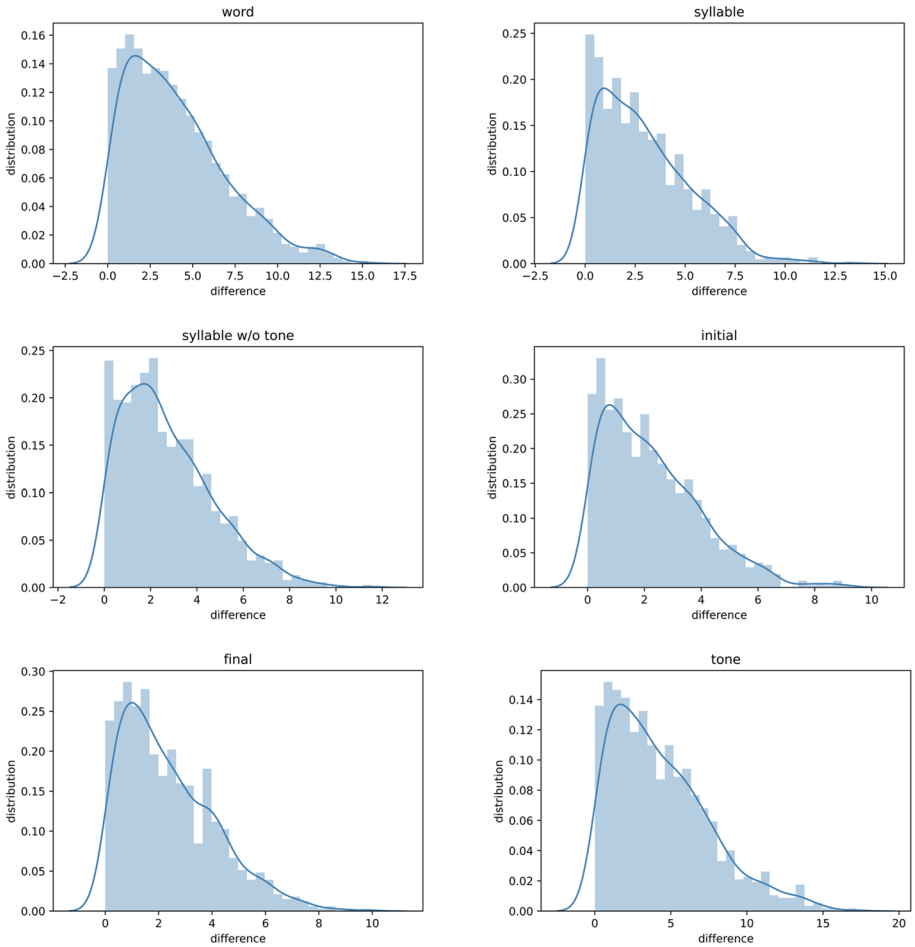

Fig. 6 Fisher permutation test of pretest–posttest improvements between diagnostic and basic feedback groups ($N = 23$ for basic and $N = 19$ for diagnostic)

p-value tests. Although the sample size is only moderate, the results are robust as consistent *p*-values have been observed across different tests with potential bias taken into account.

As the *p*-value tests provided mixed results regarding the statistical significance of the advantage of diagnostic feedback over basic feedback when conducted individually on the itemized Pinyin knowledge learning, a multivariate analysis of variances (MANOVA) is desired for a more general conclusion to be made. Here, the Hotelling *t*-squared test is chosen for the group-wise comparison of the means of the score improvements in the two feedback modes in Pinyin Tutor training. The normality of the data was confirmed by a formal multivariate normality test, with the Shapiro–Wilk statistics and the corresponding *p*-values being 0.9939 and 0.3982 for pretest to posttest improvement and 0.9797 and 0.0536 for pretest to delayed posttest improvement. In addition, the assumption of the equality of covariance matrices was met with the *p*-values greater than 0.10 for the Box' M tests for both the pretest-posttest and the pretest-delayed posttest comparisons (Huberty, 2005). As shown in Table 4, the Hotelling *p*-value of 0.5951 is not significant when testing the combined advantageous effect of diagnostic feedback in the syllable without tone, initial and final levels of Pinyin knowledge. However, the *p*-value

Table 4 Pretest–posttest Hotelling statistics based on part of and full components of Pinyin knowledge ($N=23$ for basic and $N=19$ for diagnostic)

	Hotelling statistic	Hotelling *p*-value
Syllable w/o tone, initial, and final combined	0.6381	0.5951
Word, syllable, and tone combined	3.7306	0.0191
All components combined	2.3577	0.0512

of 0.0191 is significant enough when testing the combined advantageous effect of diagnostic feedback in the word, syllable, and tone levels of Pinyin knowledge. Overall, the Hotelling *p*-value of 0.0512 is significant, at about 95% confidence level, when testing the combined advantageous effect of diagnostic feedback over basic feedback in the score improvement in all the six components of Pinyin knowledge considered altogether.

To summarize, the advantageous effects of diagnostic feedback over basic feedback in the Pinyin training for score improvement from the pretest to the posttest varied across different components of Pinyin knowledge. Notwithstanding, when taking all aspects of Pinyin knowledge, diagnostic feedback's overall advantageous effect was shown as significant at a confidence level of around 95%, as confirmed by an analysis of Hotelling statistics in a MANOVA.

The results for the learners' performance difference between the pretest and delayed posttest under two training modes of feedback, with sample size $N=13$ for the basic feedback group and $N=8$ for the diagnostic feedback group, were demonstrated in Table 5. In consistency with the results of the pre-posttests comparison, the score improvement under the diagnostic feedback condition outperformed that under the basic feedback condition for all the components of Pinyin knowledge items, although not all the differences are significant enough. Specifically, the mean score improvement from the pretest to the delayed posttest in the syllable level of Pinyin was 41.85 under diagnostic feedback mode, compared to 33.86 under basic feedback mode, representing a difference of 23.6%. This corresponds to a two-sided *p*-value of 0.162 for ANOVA, a *p*-value of 0.296 for the Fisher permutation test with 1000 iterations, as shown in Fig. 7, and a *p*-value of 0.331 for the rank-sum test, which are not significant at a confidence level of 90%. However, the mean score improvement in syllable without tone was 45.20 under diagnostic feedback mode and 28.95 under basic feedback mode, which corresponds to a *p*-value of 0.006 for ANOVA, a *p*-value of 0.013 for the Fisher permutation

Table 5 Pretest-delayed posttest improvements for the two feedback conditions ($N=13$ for basic and $N=8$ for diagnostic)

	Diagnostic		Basic		Significance	
	Mean	SD	Mean	SD	Cohen's *d*	*p*-value
Word	45.1107	11.2514	38.8158	9.6220	1.3117	0.1026
Syllable	41.8540	20.2622	33.8622	11.6519	1.0112	0.1623
Syllable w/o tone	45.1956	14.2068	28.9474	10.2567	2.8045	0.0057
Initial	20.4412	6.1668	12.7554	6.9615	2.6432	0.0080
Final	32.9525	8.7062	19.9935	3.7862	3.9556	0.0004
Tone	28.1391	29.3215	14.7757	6.9898	1.2556	0.1122

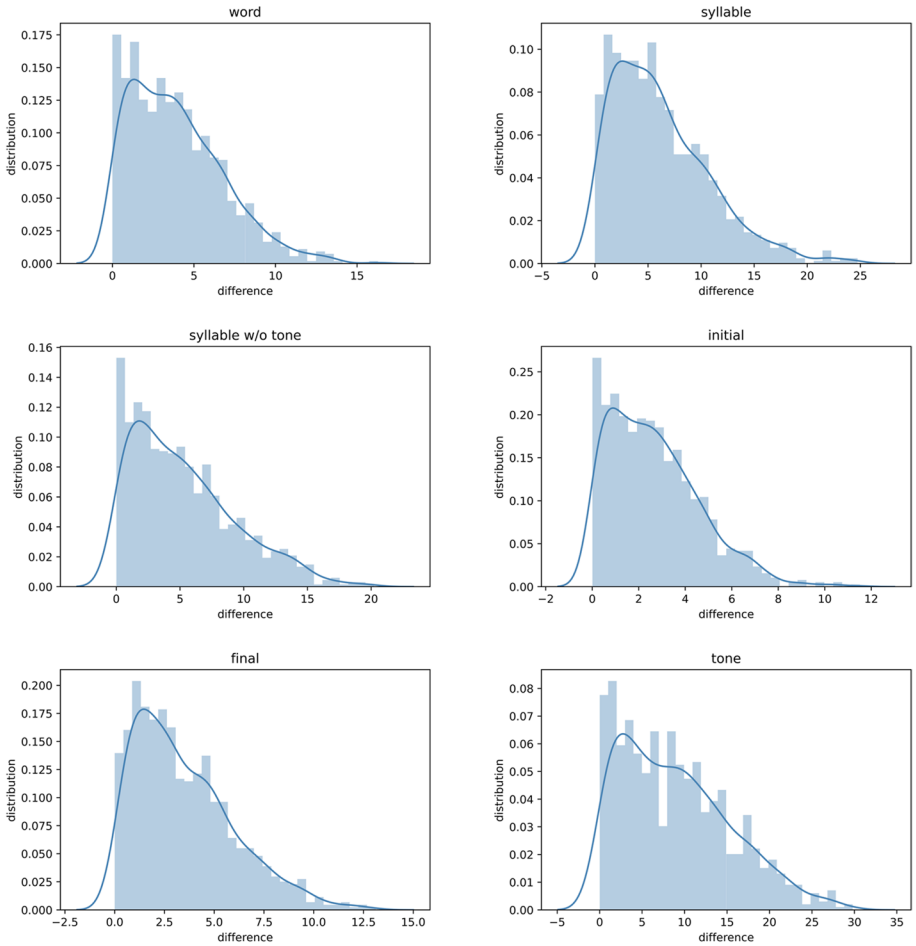

Fig. 7 Fisher permutation test of pretest-delayed posttest improvements between diagnostic and basic feedback groups ($N = 13$ for basic and $N = 8$ for diagnostic)

test, and a p-value of 0.025 for the rank-sum test, each of which is significant at a confidence level of 95%. The comparative barplots of the three kinds of p-values for assessing to what extent diagnostic feedback had outperformed basic feedback in acquiring each component of Pinyin knowledge are presented in Fig. 8. The learning-enhancing effect of diagnostic feedback is significantly higher than that of basic feedback in the syllable without tone, initial, and final levels of Pinyin knowledge while still marginally higher in the other three aspects of Pinyin knowledge, namely, word, syllable, and tone.

Similarly, the Hotelling t-squared test was conducted for the group-wise comparison of the means of the score improvement to assess the overall advantage of diagnostic feedback over basic feedback in Pinyin training. As shown in Table 6, the Hotelling p-value is not significant for the combined advantageous effect of diagnostic feedback in word, syllable, and tone levels of Pinyin knowledge. Nevertheless, the Hotelling p-value of 0.0138 is significant at about 99% confidence level when testing the combined advantageous effect of diagnostic feedback in the syllable without tone, initial, and final levels of Pinyin knowledge. The overall Hotelling p-value of 0.0874 is still significant at a 90% confidence level when testing the combined advantageous effect of diagnostic feedback in all the components of Pinyin knowledge under consideration.

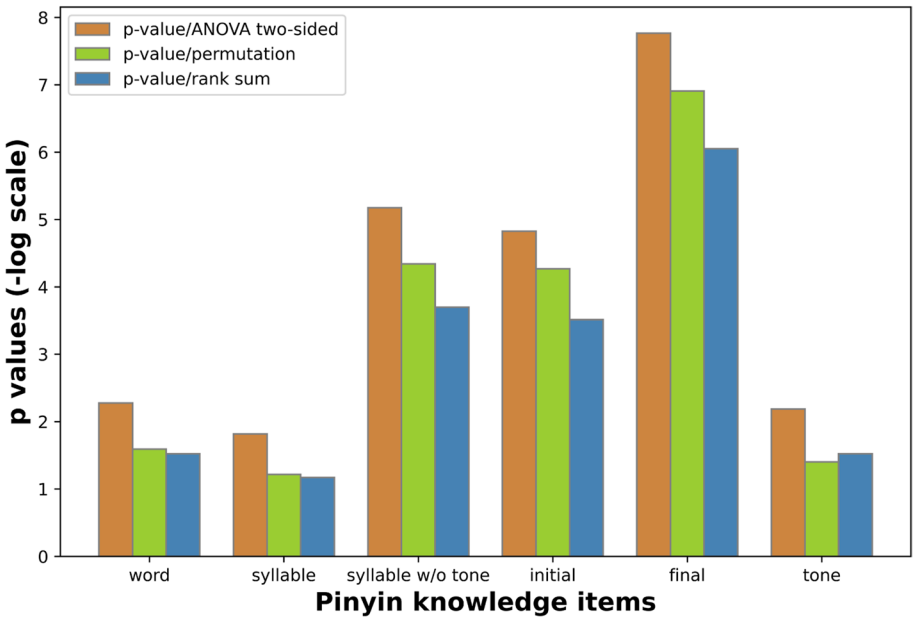

Fig. 8 Significance of pretest-delayed posttest improvements for the two feedback conditions ($N = 13$ for basic and $N = 8$ for diagnostic)

Table 6 Pretest–posttest Hotelling statistics based on part of and full components of Pinyin knowledge ($N = 13$ for basic and $N = 8$ for diagnostic)

	Hotelling statistic	Hotelling *p*-value
Syllable w/o tone, initial, and final combined	0.5451	0.6581
Word, syllable, and tone combined	4.7586	0.0138
All components combined	2.3567	0.0874

To summarize, when taking all aspects of Pinyin knowledge as a whole, the overall combined advantageous effect of diagnostic feedback is significant at a confidence level of 90%, despite that such advantageous effects showed varied magnitudes across different components of Pinyin knowledge. In consolidation of the results for the score improvement between pre-posttests and that between pre- and delayed posttests, the overall learning enhancing effect of diagnostic feedback is demonstrated as stronger than that of basic feedback in Pinyin learning through the Pinyin Tutor. Nevertheless, caution should be made to generalize the nuanced component-specific pattern in consideration of the high dimensionality of the experiment and relatively small sample size, which may have resulted in a loss of statistical power. Potential noise factors could also have exerted a confounding influence on the overall empirical significance of the test.

Discussion

The extent to which diagnostic feedback proved more effective than basic feedback varied across different aspects of Pinyin and across posttest vs. delayed posttest. As elaborated in the "Results and analysis" section, diagnostic feedback demonstrated its strongest advantage over basic feedback for the levels of the word, syllable, and tone, but it had a lesser effect for syllable without tone, initial, and final. One possible explanation of such a discrepancy is that tone perception involves a smaller number of contrasts than segmental perception. The difficulty caused by Pinyin initials and finals is more delicate and multidimensional. There are many ways in which negative L1 transfers can arise. For example, it may take years of exposure to learn the correct perception and articulation of Chinese /y/ and /ü/ by some L2 learners of Chinese, in a similar challenge for a speaker of Japanese to learn English /r/ and /l/ (MacWhinney, 2015; Ingvalson et al., 2012).

The fact that these advantages for diagnostic feedback did not carry over to the delayed posttest could be related to the fact that only the most motivated students continued with the delayed posttest. Also, the average score of those delayed posttest participants in the pretest was about 9 points higher, on average, for all Pinyin knowledge assessed. Thus, it is reasonable to assume that the learners who participated in the delayed posttest had higher prior knowledge of Pinyin overall and were more highly motivated to Pinyin learning. The already relatively higher proficiency in Pinyin tone made it harder for them to increase further their test scores in this respect through repeated Pinyin training with diagnostic feedback. However, diagnostic feedback did provide them with more advantages to maneuver the more challenging part of Pinyin, such as some subtle initials, finals, and their delicate pronounceable combinations.

One noteworthy remark is that the correlations between the learners' score improvement and the number of times the sound contrast was listened to are insignificant. For instance, such correlation calculated for the word dimension improvement in the diagnostic group from pretest to posttest is only −0.039 with a p-value of 0.861. One explanation of why repeated sound contrast listening did not seem to contribute to the dictation skill improvement comes from the results of the error analysis, which showed that 6% of the errors with initials and more than 10% of the errors with finals were produced due to orthographic transfer from English or other L1s. Some pronounceable sounds in Pinyin may not have a clear mapping in the corresponding L1 phonological categories. The current feedback designed for phonological errors did not consider the misconceptions of L1 orthographic transfer. Also, the sound comparison might have presented too many contrasting features simultaneously, so the learner's working memory was overloaded (Clark et al., 2006). These said, more finely tuned experiments are needed towards a more definite understanding of the concerned correlations, as there could be many reasons for replaying the contrasts or not, including possibly pure technical reasons.

The feedback in the current study primarily focused on learning the Pinyin knowledge of the Chinese language. However, the pedagogical benefit of such a feedback-embedded CALL environment is extendable to a broad spectrum of SLA in general. Given other equal conditions, whether an educational platform can provide intelligent feedback is emblematic of whether individual variabilities have been finely taken care of throughout the learning process, which is essential to a successful SLA (MacWhinney, 2021; Pawlak,

2022). In addition, there has been tremendous development in automated writing assessment. Integrating feedback and CALL has helped foster self-efficacy and self-regulated learning from the learner's perspective and decreased the cost and time spent in assessment from the educator's perspective. For instance, Zhang and Wu (2021) demonstrated that Chinese learners' speaking proficiency could be effectively scored by lexical richness indices profiled by the learner's speech discourse. The lexical richness measures D, LogTTR, and RootTTR have been shown as particularly effective in assessing the level of nativeness of the L2 learners' speech production of Chinese. Given the fallibility and surging cost of human raters, the automated feedback method employed in the current study should be particularly useful for L2 educators who strive for a reliable automated rating tool based on the speech corpus of the learners. Facilitated primarily by the ever-increasing machine learning and deep learning technologies, automated assessment has already been extensively adopted for the English language (Deane, 2013). Incorporating such automated language assessment with intelligent diagnostic feedback is advisable to continuously and dynamically foster language teaching and learning more effectively.

Conclusion

Language learning is complex. Such complexity is epitomized in the UCM framework by the emergent nature of language acquisition, the competition between various forms or patterns at different timeframes, and the multidimensionalness of the structures and processes involved. As an implication, the fundamental principle of language instruction should be to optimally foster various protective factors while overcoming or harnessing various risk factors at each stage of language development, where the major risk factors, in the context of L2 learning, include entrenchment, misconnection, and negative transfer. While there are many possible ways to implement these principles, feedback has been shown to be one of the most critical learning and teaching strategies for L2 acquisition, and its importance is further emphasized in a technology-assisted language learning environment, where, devoid of face-to-face coaching that a classroom-instruction may offer, the errors in learning are prone to be entrenched. The current study undertakes to examine how feedback in general, and the quality of feedback in terms of whether learners' mistakes are timely and delicately diagnosed, may facilitate the acquisition of Chinese Pinyin knowledge among the L2 learner of Chinese under an innovative CALL environment.

The findings showed that repeated training with feedback significantly enhanced Pinyin learning as measured by the score improvements between the tests arranged before and after the training sessions. Statistically significant improvement was demonstrated in each component of Pinyin skills being trained and tested, namely, word, syllable, syllable without tone, initial, final, and tone. High retention of Pinyin knowledge acquired through the Pinyin Tutor training was confirmed by the current study as a delayed posttest had demonstrated largely equivalent statistical significance in terms of accuracy rate in the Pinyin task trials by the participants, where the delayed posttest was conducted 3 months after the posttest. In juxtaposition with the findings reported by Zhang and MacWhinney (2023), the current study provides further evidence of the pedagogical effectiveness of the Pinyin Tutor, where the technology was finely embedded

with psycholinguistic theories as well as corpus-based research on the speech production of L2 learners.

As confirmed by the current study, the content of feedback, both in breadth and depth, does significantly impact the learning performance of the CALL. Specifically, diagnostic feedback led to greatere imporvement than basic feedback for the word, syllable, and tone aspects of Pinyin knowledge. The depth of feedback and how such diagnostics are integrated with the Pinyin Tutor can be further improved to more fully accommodate the multifaceted complexity of L2 learning. For example, the contrastive sound analysis may overlook the diversity and complexity of the languages that the L2 learner of Chinese may grow up with: the parents could speak different languages in the household, the neighborhood could be a melting pot of dialects, the learner could be multilingual without a clearcut L1 identity, and all such scenarios are becoming increasingly less uncommon with the fast-paced societal changes. Though challenging, the task of differentiating the wide spectrum of L1 and providing scenario-fitted feedback should be a promising direction for upgrading the Pinyin Tutor. Another aspect hindering the best performance of the current Pinyin Tutor is that many contrasting features are presented simultaneously and repeated uniformly to the learner regardless of the number of attempts the learner has made on the task item. An improvement is desirable to take into consideration the influence of working memory capacity (Goo, 2012) to avoid overfeedback at a particular stage of learning.

For the hardware and software aspect of the Pinyin Tutor, there are abundant possibilities to improve further. More gamification and social interaction features could be added to the system to meet the life and learning styles of the young generations of language learners. So far, the data analysis functions of the Pinyin Tutor are primarily descriptive and static. Although a class-wise summary or an overview of a learner's performance is available, individual learners' characteristics are not sufficiently analyzed or utilized. More sophisticated computing schemes from machine learning and artificial intelligence may play a higher facilitating role in the algorithm design, including estimating the learners' learning curve and predicting their learning performance and error patterns more proactively and dynamically.

Abbreviations

CALL	Computer-assisted language learning
SLA	Second language acquisition
UCM	Unified competition model
CALL	Computer-assisted language learning
L2	Second language
L1	First language
NLM	Native language magnet
PAM	Perceptual assimilation model
SLM	Speech learning model
MANOVA	Multivariate analysis of variances

Acknowledgements
Not applicable.

Authors' contributions
The authors jointly formulated the study, designed the methodology and procedure, and jointly collected the data and carried out preliminary analysis. YZ surveyed the literature, analyzed the data, implemented the empirical analysis, and wrote the initial manuscript. BM provided critical comments on the UCM and the results. Both authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Data used for the current study are available upon reasonable request for academic researches.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 3 November 2022 Accepted: 26 March 2023
Published online: 19 July 2023

References

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. G. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). MIT Press.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O. S. Bohn (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, *17*(2), 102–118.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299–2310.

Caldwell-Harris, C., & MacWhinney, B. (2023). Age effects in second language acquisition: Expanding the emergentist account. *Brain and Language* (in press).

Clark, R. C., Nguyen, F., Sweller, J., & Baddeley, M. (2006). Efficiency in learning: Evidence-based guidelines to manage cognitive load. *Performance Improvement*, *45*(9), 46–47.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24.

Eckman, F. R. (1981). On predicting phonological difficulty in second language acquisition. *Studies in Second Language Acquisition*, *4*, 18–30.

Eckman, F. R. (1991). The structural conformity hypothesis and the acquisition of consonant clusters in the interlanguage of ESL learners. *Studies of Second Language Acquisition*, *13*, 23–41.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues* (pp. 229–273). York Press.

Flege, J. E. (2018). It's input that matters most, not age. *Bilingualism: Language and Cognition*, *21*, 919–920.

Flege, J. E., & Bohn, O. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress*. Cambridge University Press.

Fu, M., & Li, S. (2022). The effects of immediate and delayed corrective feedback on L2 development. *Studies in Second Language Acquisition*, *44*, 2–34.

Gebril, A. (2021). *Learning-oriented language assessment: Putting theory into practice*. Taylor and Francis.

Gibbon, D., Moore, R., & Winski, R. (Eds.). (1997). *Handbook of standards and resources for spoken language systems*. Cambridge University Press.

Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. *Studies in Second Language Acquisition*, *34*, 445 - 474.

Guion, S. G., Flege, J. E., Liu, S. H., & Yeni-Komshian, G. H. (2000). Age of learning effects on the duration of sentences produced in a second language. *Applied Psycholinguistics*, *21*(2), 205–228.

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277.

Hernandez, A. E., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, *9*(5), 220–225.

Hu, B. (2010). The challenges of Chinese: A preliminary study of UK learners' perceptions of difficulty. *Language Learning Journal*, *38*(1), 99–118.

Huang, B. H. (2015). A synthesis of empirical research on the linguistic outcomes of early foreign language instruction. *International Journal of Multilingualism*, *13*(3), 257–273.

Huberty, C. J. (2005). *Applied MANOVA and discriminant analysis*. Wiley.

Ingvalson, E., Holt, L., & McClelland, J. (2012). Can native Japanese listeners learnto ifferentiate /r-l/ on the basis of F3 onset frequency? *Bilingualism: Language and Cognition*, *15*, 255–274.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

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

Kowalski, J., Zhang, Y., & Gordon, G. (2014). Statistical modeling of student performance to improve Chinese dictation skills with an intelligent tutor. *Journal of Educational Data Mining*, *6*, 3–27.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, *50*, 812–822.
Kuhl, P. K. (1998). Effects of language experience on speech perception. *Journal of the Acoustical Society of America*, *103*, 29–31.
Kuhl, P. K. (2000). Language, mind, and brain: Experience alters perception. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 99–115). MIT Press.
Lado, R. (1957). *Linguistics across cultures*. University of Michigan Press.
Li. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*, 309–365.
Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, *36*, 385–408.
Li, S. (2020). What is the ideal time to provide corrective feedback? *Language Teaching*, *53*, 96–108.
Li, S., & Roshan, S. (2019). The associations between working memory and the effects of four different types of written corrective feedback. *Journal of Second Language Writing*, *45*, 1–15.
Lyster, R. (2015). The relative effectiveness of corrective feedback in classroom interaction. In N. Markee (Ed.), *The handbook of classroom discourse and interaction* (pp. 213–228). Wiley-Blackwell.
MacWhinney, B. (1987). The competition model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 249–308). Lawrence Erlbaum.
MacWhinney, B. (1991). Reply to Woodward and Markman. *Developmental Review*, *11*, 192–194.
MacWhinney, B. (2008). A unified model. In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 341–371). Lawrence Erlbaum Associates.
MacWhinney, B. (2012). The logic of the unified model. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 211–227). Routledge Taylor & Francis.
MacWhinney, B. (2015). Multidimensional SLA. In S. Eskilde & T. Cadierno (Eds.), *Usage-based perspectives on second language learning* (pp. 22–45). Oxford University Press.
MacWhinney, B. (2018). A unified model of first and second language learning. In M. Hickmann, E. Veneziano, & H. Jisa (Eds.), *Sources of variation in first language acquisition: Languages, contexts, and learners* (pp. 287–310). John Benjamins.
MacWhinney, B. (2021). The competition model: Past and future. In J. Gervain, G. Csibra, & K. Kovács (Eds.), *A life in cognition: Studies in cognitive science in honor of Csaba Pléh* (pp. 3–16). Springer.
MacWhinney, B., & Fromm, D. (2022). Language sample analysis with TalkBank: An update and review. *Frontiers in Communication*, *7*, 865498.
Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Routledge.
Munro, N., Baker, E., McGregor, K., Docking, K., & Arciuli, J. (2012). Why word learning is not fast. *Frontiers in Psychology*, *3*, 41. https://doi.org/10.3389/fpsyg.2012.00041
Patkowski, M. S. (1990). Age and accent in a second language: A reply to James Emil Flege. *Applied Linguistics*, *11*, 73–89.
Pawlak, M. (2019). Investigating language learning strategies: Prospects, pitfalls and challenges. *Language Teaching Research*, *25*, 817–835.
Pawlak, M. (2022). Research into individual differences in SLA and CALL: Looking for intersections. *Language Teaching Research Quarterly*, *31*, 200–233.
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878-912.
Qi, Z., Han, M., Garel, K., San Chen, E., & Gabrieli, J. D. E. (2015). White-matter structure in the right hemisphere predicts Mandarin Chinese learning success. *Journal of Neurolinguistics*, *33*, 14–28.
Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis: A research tool (2nd edition)*. Springer.
Stickler, U., & Shi, L. (2013). Supporting Chinese speaking skills online. *System*, *41*(1), 50–69.
Tsou, W., Wang, W., & Tzeng, Y. (2006). Applying a multimedia storytelling website in foreign language learning. *Computers and Education*, *47*(1), 17–28.
Zhang, Y. (2020). An examination of corrective, reflective, and rule-based feedback in Chinese classifier acquisition in a CALL environment. *Theory and Practice in Language Studies*, *10*, 1558–1565.
Zhang, Y., & Li, R. (2016). The role of morphological awareness in the incidental learning of Chinese characters among CSL learners. *Language Awareness*, *25*(3), 179–196.
Zhang, Y., & MacWhinney, B. (2023). The role of novelty stimuli in second language acquisition: Evidence from the optimized training by the Pinyin Tutor at TalkBank. *Smart Learning Environments*, *10*, 3. https://doi.org/10.1186/s40561-023-00223-3
Zhang, Y., & Wu, W. (2021). How effective are lexical richness measures for differentiations of vocabulary proficiency? A comprehensive examination with clustering analysis. *Language Testing in Asia*, *11*. https://doi.org/10.1186/s40468-021-00133-6.

Publisher's Note