Open Access Library Journal
2026, Volume 13, e15637
ISSN Online: 2333-9721
ISSN Print: 2333-9705

The Impact of Voice-Interactive Multimodal Instruction on Chinese Character Literacy Outcomes and Cognitive Load in Early Primary School: A Pilot Study

Wuting Huang*, Wei Zhu, Shirui Hong, Haosheng Lu, Lulu Shao

College of Education, Zhejiang Normal University, Jinhua, China
Email: *2907946389@qq.com

How to cite this paper: Huang, W.T., Zhu, W., Hong, S.R., Lu, H.S. and Shao, L.L. (2026) The Impact of Voice-Interactive Multimodal Instruction on Chinese Character Literacy Outcomes and Cognitive Load in Early Primary School: A Pilot Study. *Open Access Library Journal*, **13**: e15637.
https://doi.org/10.4236/oalib.1115637

Received: June 15, 2026
Accepted: July 19, 2026
Published: July 22, 2026

Abstract

Chinese character literacy instruction for lower primary grades is commonly confronted with practical challenges, including weak association between character form, sound and meaning, high cognitive load in single-modal learning, and poor sustained attention of students. As a pilot study, this research adopted a quasi-experimental design and recruited 44 second-grade students from two parallel classes of a primary school. The experimental group received literacy instruction via a speech-interactive multimodal system, while the control group used the traditional audio reading-aloud mode. After a 2-week teaching intervention, the effects of this teaching mode on students' literacy performance, cognitive load and learning interest were investigated. The results indicated that the total literacy score of the experimental group was significantly higher than that of the control group ($p < 0.001$), with differentiated performance across question types: the improvement in character pronunciation recognition was extremely significant, whereas no significant improvement was found in character form discrimination. The total cognitive load of the experimental group was significantly lower than that of the control group ($p < 0.001$); dimensional analysis revealed that extraneous cognitive load decreased remarkably, while germane cognitive load did not increase significantly. The total learning interest score of the experimental group was significantly higher than that of the control group ($p < 0.01$). However, stratified analysis by academic level showed that learning interest improved significantly in the high-score group, while no significant gain was observed in the low-score group. This pilot study preliminarily indicates that, compared with the traditional audio reading-aloud mode, speech-interactive multimodal teaching can strengthen the form-sound association and reduce extraneous cognitive load in the short term. Nevertheless, it has limitations in character form processing, deep meaning construction and motivation protection for low-achieving students, which requires supporting instructional design to compensate.

Subject Areas

Education, Educational Technology, Primary Education

Keywords

Multimodal Interaction, Voice Interaction, Chinese Character Literacy, Cognitive Load, Lower Primary School

1. Introduction

1.1. Research Background: Practical Challenges in Primary School Literacy Teaching

Chinese character literacy is a core task of language teaching in the lower primary years and serves as a vital foundation for the subsequent development of pupils' reading and writing abilities. Chinese characters possess the triple attributes of form, sound and meaning; the essence of literacy lies in establishing a robust mental connection between these three elements. However, current literacy instruction in the lower primary years remains predominantly traditional, with teachers presenting character forms via textbooks and blackboard writing, conveying pronunciation through model reading, and reinforcing memory through mechanical copying. This modally segregated approach—involving visual observation of form, auditory perception of sound and manual copying—presents three prominent issues.

Firstly, the link between form, sound and meaning is weak. Information input through a single channel struggles to form multisensory encoding, leading to students frequently recognising characters but mispronouncing them, or being able to pronounce them but failing to understand their meaning; the rate of confusion regarding polysyllabic and homophonic characters remains persistently high. Secondly, there is an imbalance in cognitive load. Younger pupils have limited working memory capacity. The linear delivery of static text and images alongside one-way audio forces them to process multiple tasks—character recognition, phonetic memorisation and semantic comprehension—within a short timeframe. This results in a high external cognitive load, which diverts cognitive resources away from the construction of deeper meaning. Thirdly, a decline in learning motivation. The one-way, rote-learning approach of repetition and copying lacks immediate feedback and interactivity. Students' attention spans are short, making them prone to boredom and frustration, and resulting in insufficient intrinsic motivation.

With the advancement of educational informatisation, multimodal interactive technologies offer new possibilities for overcoming the aforementioned challenges. Among these, voice interaction technology, with its features of natural conversation, real-time assessment and personalised feedback, aligns closely with the phonological encoding requirements of Chinese character literacy. Integrating voice interaction into literacy instruction to establish a multi-channel closed loop encompassing visual, auditory and oral expression holds promise for breaking through the limitations of traditional single-modal teaching and providing a new pathway for literacy instruction in the early years.

1.2. Research Gaps

Existing research on multimodal literacy instruction has achieved certain results, but three significant shortcomings remain. Firstly, research subjects have focused on passive multimedia. Most studies focus on pre-set multimedia materials such as animated voiceovers and image-audio combinations, which essentially remain one-way information delivery. They lack a process where students actively initiate interaction and receive immediate feedback, and there is little research on voice-based tools capable of two-way interaction. A review by Olmanson and Liu [1] points out that existing literacy-related educational technology is largely concentrated on character visualisation and mechanical drills, whilst empirical research on voice interaction remains relatively scattered. Secondly, research dimensions are biased towards academic performance and lack a differentiated perspective. Existing findings predominantly use literacy accuracy and retention rates as core indicators, with few studies simultaneously examining cognitive load and learning interest; even when cognitive load is addressed, the focus is largely on total scores, with little distinction made between the differentiated impacts of intrinsic, extrinsic, and related loads, making it difficult to reveal the true mechanisms of action. Thirdly, there is a shortage of lightweight solutions for localised empirical research, and insufficient attention is paid to group differences. Domestic studies often rely on expensive specialised teaching equipment, limiting their universality; furthermore, they rarely examine differences in outcomes among students of varying academic levels, tending to assume that the technology is equally effective for all students and overlooking the potential risk of frustration for those with weaker foundations. Recent systematic reviews also note that empirical research on voice-interactive literacy tools for young Chinese learners remains limited, especially studies focusing on cognitive mechanisms.

Consequently, this study employs a lightweight, voice-interactive multimodal literacy system as an intervention tool, focusing on the literacy teaching context of Year 2 primary school pupils. By examining both overall scores and specific dimensions, it preliminarily explores the system's impact on pupils' literacy outcomes, cognitive load and learning interest, thereby providing more targeted practical guidance for frontline teaching.

1.3. Research Questions and Hypotheses

Research Questions

1) Compared to traditional audio-based repetition teaching, what impact does voice-interactive multimodal teaching have on the overall literacy scores and scores by question type of lower-year primary school pupils?

2) What effects does voice-interactive multimodal instruction have on students' overall cognitive load and the three specific sub-dimensions of cognitive load?

3) Are there differences in the impact of voice-interactive multimodal teaching on the learning interest of students with different academic abilities?

Research Hypotheses

H1: Students in the experimental group will achieve significantly higher overall literacy scores than those in the control group; the greatest improvement will be in phonetic recognition, whilst the smallest improvement will be in character recognition. H2: Students in the experimental group will exhibit significantly lower total cognitive load scores than those in the control group; specifically, extrinsic load will decrease significantly, whilst related load will show no significant change. H3: Students in the experimental group will exhibit significantly higher total learning interest scores than those in the control group; interest will increase significantly in the high-achieving group, whilst no significant increase will be observed in the low-achieving group.

1.4. Research Significance

At the theoretical level, this study extends multimedia learning theory and cognitive load theory to the context of voice-interactive literacy learning. Through detailed, multi-dimensional and stratified analysis, it reveals the boundaries of voice interaction's impact on literacy learning and the differences in its mechanisms, thereby enriching the cognitive research perspective on primary school Chinese character and vocabulary instruction. At the practical level, the lightweight voice-based literacy teaching programme proposed in this study is simple to implement and cost-effective. Furthermore, by clarifying its applicable scenarios and limitations, it provides a reference for frontline Chinese language teachers in designing tiered literacy activities, whilst also offering educational principles for the age-appropriate and differentiated design of educational tools. It should be noted that this study is a pilot study with a limited sample size and intervention duration. It aims to preliminarily validate the effectiveness of the intervention and the feasibility of the programme, laying the groundwork for subsequent large-scale empirical research; therefore, its conclusions should be interpreted with caution.

2. Literature Review and Theoretical Framework

2.1. Multimodal Learning and Primary School Chinese Character Literacy Teaching

Multimodal learning refers to a learning approach that integrates two or more sensory channels—such as visual, auditory and kinesthetic—for information processing. The cognitive theory of multimedia learning proposed by Mayer [2] posits that humans possess two relatively independent information processing channels: visual and auditory. Synergistic input through these dual channels can effectively expand working memory capacity and enhance learning outcomes. The theory's principles—including the modality principle and the redundancy principle—have been substantiated by extensive research and have become the core basis for educational multimedia design.

Chinese character literacy is, at its core, a cognitive process involving multisensory coordination. Domestic research on literacy instruction indicates that the encoding of Chinese characters' form, sound and meaning relies on distinct sensory channels: character recognition relies on the visual channel, phonetic processing on the auditory and speech-motor channels, whilst comprehension of meaning depends on the multi-channel integration of context and experience. Traditional teaching methods, such as purely visual memorisation or purely auditory repetition, sever the connections between these channels, resulting in single-channel encoding and rapid forgetting.

Voice interaction is an emerging modality in the field of language learning in recent years. Relevant research indicates that speech recognition and real-time assessment technologies can effectively enhance learners' phonological awareness and pronunciation accuracy, and have already demonstrated positive results in the teaching of pinyin pronunciation and Chinese character reading. Research by Otsu and Izumi [3] confirms that voice prompts can facilitate active recall of Chinese character forms, offering greater processing depth compared to purely visual presentation. Its core advantage lies in two-way interaction and immediate feedback; students are no longer passive recipients of audio but actively speak out, receiving instant feedback on correctness and corrective guidance, thereby forming a cognitive loop of input, output, feedback and correction. In addition, generative AI-based multimodal resource construction [4] has also become a new development direction in this field. However, empirical research on the application of voice interaction in primary school Chinese character literacy instruction remains scarce, and few studies have examined its differential impact on students with varying abilities and across different question types.

2.2. Cognitive Load Theory and Primary School Literacy Learning

Cognitive load theory was proposed by Sweller [5]. This theory posits that human working memory capacity is limited, and instructional design should aim to minimise extrinsic cognitive load whilst maximising relevant cognitive load. Cognitive load is categorised into three types: intrinsic load, determined by the complexity of the learning material itself; extrinsic load, caused by inappropriate presentation of information and optimised through instructional design; and relevant load, which represents the active cognitive engagement required to facilitate schema construction. In their seminal review, Paas *et al.* [6] systematically examined the interrelationships between these three types of load, establishing the theoretical framework for the application of cognitive load in instructional design.

For pupils in the lower years of primary school, the intrinsic load of literacy learning is inherently high. Chinese characters are logographic, with complex structures and a multitude of homophones and polysyllabic words; moreover, pupils in the lower years have not yet fully developed phonological awareness or character perception. In traditional teaching, the use of one-way audio playback and static text-image presentations further increases external load; pupils must independently verify the correctness of pronunciations and repeatedly recall character forms, with a significant amount of cognitive resources being consumed in inefficient self-checking, thereby crowding out resources intended for the construction of connections between form, sound and meaning.

A well-designed multimodal approach can optimise cognitive load through dual-channel sharing and cognitive unloading. On the one hand, the parallel processing of auditory and visual channels can alleviate the information pressure on the single visual channel; on the other hand, immediate speech assessment and correction replace students' self-assessment and repeated trial-and-error, thereby reducing ineffective cognitive expenditure. However, it should be noted that more modalities are not necessarily better; if information becomes redundant or functions overly complex, this may actually increase external load, thereby defeating the original design intent.

It is worth noting that existing research largely assumes that reducing extrinsic load inevitably enhances learning outcomes. However, in reality, whether the released cognitive resources can be converted into relevant loads for deep meaning construction depends on the learner's metacognitive abilities and task guidance. As younger students have limited metacognitive levels, reducing cognitive load does not necessarily equate to deep learning; this hypothesis still requires empirical testing. Regarding measurement tools, the Differentiated Cognitive Load Questionnaire developed by Klepsch and Seufert [7] comprehensively distinguishes between the three types of load, but the original version was designed for adult populations; Altmeyer *et al.* [8] validated a child-adapted version of the Differentiated Cognitive Load measurement scheme, confirming that primary school pupils in the lower and middle years can effectively distinguish between intrinsic and extrinsic loads, though the validity of the measurement of related loads still requires further verification. Recent validation work on differentiated cognitive load measurement [9] further supports the feasibility of differentiated cognitive load measurement in lower primary school pupils, though instrument refinement remains an ongoing direction. There is currently a lack of fully validated, mature tools for three-dimensional differentiated measurement in younger children, necessitating adaptive revisions based on adult scales.

2.3. Integration of the Theoretical Framework

This study integrates the dual-channel hypothesis of multimedia learning with cognitive load theory to construct its theoretical framework. Voice-interactive multimodal literacy instruction influences learning outcomes through three pathways:

Dual-channel encoding pathway: visual presentation of character form and meaning, auditory transmission of standard pronunciation, and oral output reinforcing phonological memory. Multi-channel collaborative encoding establishes connections between form, sound and meaning, with greater gains observed in the phonological dimension than in the visual dimension;

Cognitive Load Optimisation Pathway: Immediate feedback reduces the external load of self-correction for students; however, whether the associated cognitive load is simultaneously reduced depends on instructional guidance and the students' own abilities;

Affective Motivation Pathway: Interactivity and positive feedback enhance learning interest and self-efficacy; however, this effect is subject to an academic threshold and may have a counterproductive effect on students with weaker foundations.

The combined action of these three pathways ultimately results in an overall positive yet internally differentiated learning outcome pattern. It should be noted that this study primarily examines the overall intervention effect and has not conducted independent experimental separation of the three pathways; therefore, subsequent discussions will combine quantitative results with qualitative data to conduct an inferential analysis of the relative contributions of the three pathways rather than causal verification.

3. Research Methodology

3.1. Research Design

This study employed a quasi-experimental design, selecting two parallel Year 2 classes from a primary school to serve as the experimental and control groups respectively. The experimental group received literacy instruction using a voice-interactive multimodal literacy system, whilst the control group was taught using a traditional audio-repetition method. The teaching content, duration and teachers were identical for both groups; only the teaching tools and interaction methods differed. The study comprised three phases: a pre-test, an intervention and a post-test, supplemented by semi-structured interviews with pupils to gather qualitative data.

This study is positioned as a pilot study, aiming to test the feasibility and preliminary effectiveness of the intervention programme, and to provide a methodological and data foundation for subsequent large-scale research.

3.2. Research Participants

This study adopted a cluster random assignment approach: two intact parallel Year 2 classes from a public primary school in Jinhua, Zhejiang Province were selected and randomly assigned to the experimental group and the control group, with 22 pupils in each group (N = 44 in total). All pupils had no diagnosed reading difficulties or hearing impairments, and had no prior experience with similar voice-interactive literacy tools.

To ensure the internal validity of the quasi-experimental design, we verified homogeneity at both the individual and class levels. At the individual level, independent samples t-tests showed no significant difference in pre-test literacy scores ($t = 0.263$, $p = 0.794$) or final Chinese language grades ($p > 0.05$) between the two groups. At the class level, the two classes had identical class size, a consistent gender ratio, and were taught by the same Chinese language teacher with equivalent teaching experience and proficiency. These checks confirmed that the two groups were homogeneous and comparable at baseline.

This study strictly adhered to ethical guidelines for research involving children. Prior to implementation, consent was obtained from the school and the teaching staff; informed consent forms were distributed to all parents and signed; all student data were anonymised and used solely for research analysis; the intervention content strictly aligned with the curriculum progression and did not impose any additional academic burden on the students.

3.3. Research Tools

Intervention Tool: Voice-Interactive Multimodal Literacy System

The literacy system used in this study was jointly developed by the project team. Designed for literacy learning scenarios in the lower years of primary school, it can run on standard terminal devices and supports offline use. The core teaching functions of the system are as follows:

1) Text recognition and model reading function: When students point the device at new characters in the textbook, the system automatically identifies the text's position and simultaneously outputs the standard Mandarin pronunciation and a concise definition;

2) Speech repetition and real-time assessment: When students repeat new characters and words, the system recognises their pronunciation in real time and assesses accuracy, highlighting syllables that are mispronounced or pronounced incorrectly, and providing a demonstration of the correct pronunciation;

3) Contextual correction for homophones: For commonly confused homophones in the text, the system automatically determines the correct pronunciation based on the surrounding context and provides targeted guidance on pronunciation rules;

4) Interactive voice-based extension: Students can ask questions about new characters via voice input, and the system generates engaging answers tailored to the cognitive level of younger pupils.

The core feature distinguishing this system from traditional audio tools is its two-way interaction and instant feedback. The system is developed on a domestically produced software and hardware platform and supports offline operation; specific technical implementation details fall outside the scope of this educational research discussion.

Assessment Tools

1) Chinese Character Literacy Test Paper: Based on the teaching syllabus for the

second-year Chinese language textbook (second half) published by People's Education Press, a test paper comprising 20 target characters was compiled. The paper is marked out of 100 and consists of three question types: character pronunciation recognition (selecting the correct pinyin for a given character), worth 40 marks; character form analysis (filling in the blanks with the correct character), worth 30 marks; and character meaning matching (matching words with pictures), worth 30 marks. The test paper was reviewed for content validity by two primary school Chinese language teachers working in the field. The Cronbach's α coefficient for internal consistency was calculated at 0.82, indicating good reliability.

2) Children's Cognitive Load Scale: This study is based on the three-dimensional differentiated cognitive load questionnaire proposed by Klepsch and Seufert [7]. Drawing on the principles for child-friendly adaptation outlined by Altmeyer *et al.* [8], the original items were simplified linguistically and adapted to suit the context, resulting in a cognitive load scale suitable for lower primary school pupils. The scale comprises three dimensions—internal load, external load and related load—consisting of nine items in total. It employs a 5-point rating scale, where 1 indicates 'strongly disagree' and 5 indicates 'strongly agree'; a higher score indicates a higher level of cognitive load on the corresponding dimension. The intrinsic load dimension measures the difficulty of the learning material as perceived by the student; the extrinsic load dimension measures the additional burden arising from the manner in which information is presented; and the related load dimension measures the degree of active effort the student invests in meaning-making. In the pilot study, the internal consistency coefficients for each dimension of this scale were as follows: Internal load Cronbach's $\alpha = 0.76$, external load Cronbach's $\alpha = 0.78$, related load Cronbach's $\alpha = 0.69$, and total scale Cronbach's $\alpha = 0.79$. The reliability of the related load dimension is slightly lower; the causes of this and the measurement limitations will be analysed in the discussion section.

3) Literacy Learning Interest Questionnaire: This questionnaire was adapted from mature academic emotion scales for primary school Chinese learning used in existing domestic empirical studies. We first extracted 8 candidate items related to learning interest from the established item pool, then simplified the wording to match the cognitive and language level of lower primary school pupils. After content validity evaluation by two frontline Chinese teachers with more than 5 years of lower-grade teaching experience, 4 items with the highest relevance and clarity were retained, covering learning enjoyment, active learning willingness, learning pleasure perception, and learning engagement intention.

A 5-point Likert scale was used (1 = strongly disagree, 5 = strongly agree), with higher scores indicating greater interest in literacy learning. The item-level content validity index (I-CVI) of all four items was 1.00, and the scale-level average content validity index (S-CVI/Ave) was 1.00, indicating excellent content validity. The Cronbach's α coefficient of the questionnaire in this study was 0.81, demonstrating good internal consistency.

4) Semi-structured interview guide: Following the conclusion of the interven-

tion, four students were randomly selected from each of the high- and low-scoring groups within the experimental group for brief one-to-one interviews. Key questions included: 'Do you like using this tool to learn new characters?', 'Do you feel it helps you remember new characters more easily?', and 'What do you like or dislike most about it?'. Each interview lasted approximately five minutes and was fully audio-recorded and transcribed.

3.4. Research Procedure

The entire study spanned three weeks, with the specific procedure outlined below.

Week 1: Pre-test phase One week prior to the intervention, both groups of students underwent pre-tests for character recognition, cognitive load, and learning interest. The tests were conducted uniformly during Chinese language lessons and lasted approximately 20 minutes.

Weeks 2-3: Intervention Phase A two-week teaching intervention was conducted, comprising three sessions per week, each lasting 20 minutes, scheduled during morning reading sessions and after-school support periods. The teaching content consisted of new vocabulary from Unit 4 of the People's Education Press Year 2, Volume 2 textbook; the learning content was identical for both groups.

1) Teaching procedure for the experimental group: Step 1: Recognition and model reading. Students point to the new characters in the textbook; the system recognises them and plays the standard pronunciation, allowing students to gain an initial understanding of the sounds and forms of the characters; Step 2: Repetition and assessment. Students follow the system to read along character by character, with the system providing real-time feedback on the accuracy of their pronunciation; Step 3: Correction and consolidation. For incorrectly pronounced characters, the system repeats the model reading and highlights key pronunciation points, with students repeating the reading until correct; Step 4: Interactive extension. Students can ask questions via voice input, and the system provides engaging extension activities, with the teacher providing supplementary explanations in real time.

2) Control group teaching process: A traditional audio-repetition model was adopted, playing standard recordings of the same new characters and words. Students repeated the pronunciation as a group and practised freely, with the teacher providing unified correction and explanation. The duration and content of the lessons were identical to those of the experimental group.

During the intervention period, researchers observed and recorded classroom conditions in real time to ensure that the teaching procedures for both groups were strictly controlled.

Week 3: Post-intervention phase On the day following the conclusion of the intervention, both groups simultaneously underwent post-intervention assessments of character recognition, cognitive load and learning interest. Subsequently, semi-structured interviews were conducted with eight students selected from the experimental group.

3.5. Data Analysis Methods

Quantitative data were analysed using SPSS 26.0. Descriptive statistics (means and standard deviations) were calculated for all indicators in both groups. Paired samples t-tests were used to examine within-group changes from pre-test to post-test, and independent samples t-tests were used to compare post-test performance and gain scores between the two groups. Stratified analysis was further conducted for the learning interest variable to explore group differences across academic levels.

Given that the pre-test scores of the two groups were not significantly different at baseline, and this study is a small-sample pilot study, we adopted the post-test comparison and gain score analysis approach rather than analysis of covariance (ANCOVA). This approach is widely used in pilot studies with homogeneous baseline groups and small sample sizes. For future large-scale studies, we recommend using ANCOVA to statistically control for pre-test scores for more robust estimation.

After transcribing the qualitative interview data into text, the data was analysed using the thematic analysis method proposed by Braun and Clarke [10], with the following specific steps:

1) Familiarisation with the data: repeated reading of the interview transcripts to gain an initial grasp of the overall content;

2) Generate initial coding: two researchers independently performed open coding of the text, marking meaningful semantic units;

3) Identifying candidate themes: aggregating scattered codes into potential themes and clarifying the logical relationships between them;

4) Reviewing themes: verifying, layer by layer, the alignment of themes with the original data and research questions, and eliminating themes that could not be substantiated;

5) Define and name themes, clarifying the core meaning, boundaries and typical expressions of each theme;

6) Draft the analysis report, distilling final conclusions and selecting representative quotations as supporting evidence.

The coding process employed a semi-structured framework, combining the research questions with three categories of initial coding—learning experiences, cognitive perceptions and motivational feedback—whilst retaining scope for open coding to accommodate new themes emerging from the data. The coding consistency between the two independent coders was Cohen's $\kappa = 0.82$, meeting the standard for good agreement; discrepancies in coding were resolved through mutual discussion.

4. Research Findings

4.1. Comparison of Literacy Scores

Inter-group Comparison of Overall Literacy Scores

The pre- and post-test results for total literacy scores in both groups are shown in Table 1.

Table 1. Comparison of pre- and post-test total literacy scores between the two groups (M ± SD).

Group	Number of Students	Pre-test Score	Post-test score	Improvement
Experimental group	22	62.45 ± 8.32	85.36 ± 6.74	22.91
Control group	22	61.82 ± 7.95	76.59 ± 7.18	14.77

Note: Maximum score 100; improvement is calculated as the post-test mean minus the pre-test mean.

An independent samples t-test revealed no significant difference in pre-intervention literacy scores between the two groups ($t = 0.263$, $p = 0.794 > 0.05$), indicating that the two groups were homogeneous at baseline. Following the intervention, the post-test literacy scores of the experimental group were significantly higher than those of the control group, with the difference being statistically significant ($t = 4.217$, $p < 0.001$, Cohen's $d = 1.27$), indicating a large effect size. A paired t-test within groups showed that the post-test scores for both groups were significantly higher than the pre-test scores ($p < 0.001$), but the improvement in the experimental group was markedly greater.

Inter-group Comparison of Scores by Question Type

To further analyse the differential effects of voice interaction on different dimensions of literacy, independent samples t-tests were conducted on the post-test scores for the three question types, with the results shown in Table 2.

Table 2. Comparison of post-test literacy scores by question type between the two groups (M ± SD).

Question Type	Maximum Score	Post-test for experimental group	Post-test for the control group	t-value	p-value
Phonetic recognition	40	35.68 ± 3.12	29.55 ± 4.07	5.621	<0.001
Character form discrimination	30	24.27 ± 3.05	23.64 ± 2.98	0.684	0.498
Lexical matching	30	25.41 ± 2.86	23.40 ± 3.12	2.215	0.032

Note: $p < 0.05$ indicates a significant difference.

Results varied significantly across question types: In the phonetic recognition dimension, the experimental group's scores were extremely significantly higher than those of the control group, with a very large effect size, constituting the primary source of the overall score improvement; In the character shape discrimination dimension, there was no significant difference between the two groups, with the speech interaction yielding almost no additional benefit; In the character meaning matching dimension, the experimental group scored slightly higher than the control group, with the difference reaching statistical significance but the effect size being moderate.

4.2. Comparison of Cognitive Load

Inter-group Comparison of Total Cognitive Load Scores

The pre- and post-test results for the total cognitive load scores of the two groups are shown in Table 3.

Table 3. Comparison of pre- and post-test total cognitive load scores between the two groups (M ± SD).

Group	Number of Participants	Pre-test score	Post-test score
Experimental group	22	3.12 ± 0.58	2.47 ± 0.61
Control group	22	3.08 ± 0.62	3.21 ± 0.57

Note: A 5-point scale was used; a higher score indicates a higher cognitive load.

There was no significant difference in baseline cognitive load between the two groups prior to the intervention ($t = 0.221$, $p = 0.826 > 0.05$). Following the intervention, the total cognitive load score in the experimental group was significantly lower than that in the control group, with the difference being statistically significant ($t = -4.103$, $p < 0.001$, Cohen's $d = -1.24$). Within-group comparisons showed that cognitive load decreased significantly in the experimental group ($p < 0.001$), whilst there was no significant change in the control group ($p > 0.05$).

Inter-group comparisons of cognitive load by dimension

An independent samples t-test was conducted on the post-test scores for the three types of cognitive load; the results are shown in Table 4.

Table 4. Comparison of post-test cognitive load scores by dimension between the two groups of students (M ± SD).

Cognitive Load Dimension	Experimental group	Control Group	t-value	p-value
Intrinsic load	2.64 ± 0.71	2.71 ± 0.68	−0.335	0.739
External load	2.18 ± 0.65	3.25 ± 0.72	−5.176	<0.001
Correlated load	2.59 ± 0.63	2.52 ± 0.70	0.347	0.730

Note: Intrinsic load refers to the burden arising from the difficulty of the material itself; extrinsic load refers to the additional burden caused by the manner in which information is presented; and relevant load refers to the effective cognitive effort expended on meaning construction.

The results across dimensions reveal clear structural differences: In the extrinsic load dimension, the experimental group scored significantly lower than the control group, validating the load-reducing effect of voice interaction, as the system replaced the ineffective cognitive effort expended by students on self-correction; In the intrinsic load dimension, there was no significant difference between the two groups, consistent with the expectations of controlling for variables, as the learning materials were identical for both groups; In the related load dimension, there was no statistical difference between the two groups; the cognitive resources freed up by the tool did not automatically translate into active engagement by students in deep meaning-making.

4.3. Comparison of Learning Interest

Inter-group Comparison of Total Learning Interest Scores

The pre- and post-test results for the total learning interest scores of the two groups are shown in Table 5.

Table 5. Comparison of pre- and post-test total learning interest scores between the two groups (M ± SD).

Group	Number of Students	Pre-test score	Post-test score
Experimental group	22	3.25 ± 0.64	4.02 ± 0.53
Control group	22	3.31 ± 0.59	3.42 ± 0.61

Note: A 5-point scale was used; a higher score indicates greater interest in learning.

There was no significant difference in baseline learning interest between the two groups prior to the intervention ($t = -0.325$, $p = 0.747 > 0.05$). After the intervention, the experimental group's learning interest scores were significantly higher than those of the control group, with the difference being statistically significant ($t = 3.482$, $p < 0.01$, Cohen's $d = 1.05$). Within-group comparisons showed that learning interest in the experimental group increased significantly ($p < 0.001$), whilst there was no significant change in the control group ($p > 0.05$).

Stratified Comparison of Students with Different Academic Levels

We used the median of the full sample's pre-test literacy scores as the uniform cutoff point to divide students in both the experimental and control groups into high-scoring and low-scoring subgroups separately, with 11 students in each subgroup. We then compared post-test learning interest scores across subgroups; results are presented in Table 6.

It should be noted that this stratified analysis is exploratory in nature, as the sample size of each subgroup is relatively small. The results should be interpreted with caution and need to be verified in future large-sample studies.

Table 6. Comparison of post-test learning interest among students of different academic levels (M ± SD).

Group	High-achieving group	Low-scoring group
Experimental group	4.35 ± 0.42	3.68 ± 0.57
Control group	3.51 ± 0.58	3.33 ± 0.62
t-value	3.872	1.419
p-value	0.001	0.171

Note: High and low score groups were divided based on the median of pre-test literacy scores.

The stratified results reveal clear group differences: in the high-scoring group, the experimental group's interest scores were extremely significantly higher than those of the control group; the enjoyment and sense of achievement derived from voice interaction greatly stimulated the willingness to learn; in the low-scoring group, there was no statistical difference in interest between the two groups, with the experimental group only slightly higher than the control group, showing no significant improvement.

4.4. Qualitative Research Results

Four core themes were identified through thematic analysis of interview data from eight students, with representative quotations presented below to support each interpretation.

Theme 1: Immediate feedback enhances a sense of achievement in the high-scoring group

Students in the high-scoring group generally reported that they received instant correctness feedback right after reading aloud, which brought them a strong sense of accomplishment and motivated them to practice more.

- "Every time I read a character correctly, the system says 'great job' right away. I feel so happy and I want to read more and more to get more praise." (High-scoring student A)
- "I know immediately if I read it wrong, and I can fix it right away. It's better than waiting for the teacher to correct me later." (High-scoring student B)

These quotations reflect that immediate positive feedback creates a positive reinforcement cycle, where successful performance and recognition further increase students' willingness to engage.

Theme 2: Frequent corrections cause frustration in the lower-achieving group

Three students in the lower-achieving group mentioned that frequent error prompts from the system made them feel anxious and reluctant to keep reading, even when the feedback tone was friendly.

- "I always read wrong, and the system keeps telling me I'm not right. I feel nervous and I don't want to read anymore." (Low-scoring student C)
- "No matter how many times I try, I can't get it right. It makes me feel like I'm bad at reading characters." (Low-scoring student D)

These accounts indicate that repeated error reminders may undermine low-achieving students' self-efficacy, and the resulting frustration can offset the interactive appeal of the tool.

Theme 3: Interactive formats enhance overall attention span

Regardless of their performance level, students agreed that active reading and self-operated interaction was more engaging than passive listening to recordings, and helped them maintain concentration during learning.

- "I used to zone out when we just listened to the tape. But now I have to point to the character and read it out loud, so I can't wander off." (Mid-scoring student E)
- "It's more fun than just sitting and listening. I have to do it myself, so I pay more attention." (High-scoring student F)

This theme confirms that the active participation requirement of voice interaction effectively improves classroom engagement and sustained attention among younger pupils.

Theme 4: Extended features stimulate a desire to explore

Some students, especially higher-performing ones, showed strong interest in the voice-based question-and-answer extension function and were eager to explore more word combinations and background knowledge.

- "I like to ask the system what other words this character can make. It tells me interesting things I didn't know before." (High-scoring student G)

However, low-performing students rarely used this feature, as they focused most of their effort on getting the pronunciation correct.

- "I just want to get the reading right first. I don't have time to ask other questions." (Low-scoring student H)

This finding reflects the differentiated usage patterns of students with different academic foundations, and highlights the need for tiered functional design.

5. Discussion

5.1. Interpretation of Research Findings

Differentiation in Character Recognition Performance: The Limitations of Modal Functions

This study found that voice interaction significantly improved character pronunciation recognition, had almost no benefit for character shape discrimination, and produced a moderate improvement in character-meaning matching. These results corroborate the principle of modal specialisation in multimodal learning: each modality is best suited to specific cognitive processing stages, and no single modality is universally effective.

Specifically, the core of voice interaction lies in the reinforcement of auditory and speech motor pathways, which naturally align with phonetic encoding and pronunciation training; hence, the improvement in phonetic recognition is most pronounced. Conversely, the precise analysis of character forms relies on visual scanning, component decomposition and the kinesthetic sense of writing; the purely auditory modality cannot cover this processing, resulting in no difference between the two groups in character form analysis. The moderate improvement in character meaning recognition stems, on the one hand, from students' ability to better link their existing spoken language experience to character meanings once they can pronounce them accurately; on the other hand, it benefits from the system's brief explanatory extensions. However, the limited extent of this improvement indicates that the auditory modality alone struggles to support in-depth meaning construction.

This result corrects the simplistic notion that multimodal approaches are universally superior to monomodal ones, and offers guidance for classroom practice: phonetic tools should be positioned as aids for pronunciation correction, not as substitutes for visual and kinesthetic activities such as character writing and component analysis. Comprehensive literacy instruction still requires multimodal coordination.

Cognitive Load Structure: Reduced Load Does Not Equate to Deep Learning

This study found that voice interaction significantly reduced external cognitive load, but the associated cognitive load did not increase correspondingly. This finding does not fully align with the optimistic expectation that reducing external load will automatically promote deep learning, and offers important theoretical insights.

Considering the three-component structure of cognitive load, the significant reduction in external load aligns with research expectations. The system's real-time assessment and error-correction functions replaced the inefficient process of self-checking by students, directly reducing the additional cognitive expenditure associated with information presentation. It is worth noting that the subjective ratings of internal load are moderated by learners' prior knowledge levels. In this study, there was no difference in pre-test scores between the two groups, and as the same lecturer delivered the lessons, the variables of material complexity and teaching quality were objectively controlled; consequently, no systematic bias in intrinsic load was observed. Future research incorporating literacy materials of varying difficulty levels could further examine whether the load-reducing effect of voice-interactive tools remains robust under conditions of high intrinsic load.

From the perspective of cognitive load theory, the reduction in external load merely frees up cognitive resources; however, whether these resources can be converted into relevant loads for meaning construction does not occur automatically. It depends on two conditions: firstly, the learner's willingness and ability to engage in active deep processing; and secondly, clear and in-depth guidance within the teaching task. The metacognitive abilities of younger students are still developing; for most students, learning objectives remain at the superficial level of correctly pronouncing characters, and they do not actively use the saved mental energy to reflect on character meanings or make word associations. Furthermore, the basic intervention in this study primarily consisted of read-aloud assessments, lacking the design of in-depth tasks such as follow-up questioning or transfer; consequently, the released cognitive resources were not converted into relevant load.

It should be noted that the lack of significant differences in the related load dimension may also be due to limitations in the validity of the measurement tools. Related load refers to the psychological engagement involved in actively constructing meaning, which is a relatively implicit and abstract cognitive state. Younger pupils have limited capacity for understanding and self-awareness regarding such abstract concepts; they struggle to accurately distinguish between effort and deep engagement on a 5-point scale. This is the primary reason for the slightly lower reliability of this dimension in the pre-test. Insufficient measurement sensitivity may have prevented the capture of subtle, genuine changes; this limitation leaves room for improvement in future research, which should combine objective behavioural indicators with contextualised interviews to assess related load more precisely.

Differences in Interest Stratification: The Dual Nature of Immediate Feedback

This study found that the enhancement of learning interest through voice interaction is subject to a clear academic threshold: the high-scoring group showed significant improvement, whilst the low-scoring group exhibited no significant gain. This reveals the dual nature of immediate feedback: whilst it serves as a motivational tool, it may also become a source of frustration.

For students in the high-achieving group, with high accuracy rates in repetition, positive feedback predominates, creating a positive reinforcement cycle of 'answering correctly, receiving praise, and being more willing to learn', in which the enjoyment and sense of achievement derived from the interaction are fully realised. For students in the low-achieving group, however, with weak pronunciation foundations and high error rates, even when the feedback is delivered in a friendly tone, repeated error prompts constantly remind them of their own shortcomings, undermining their self-efficacy. When the sense of frustration outweighs the enjoyment, the motivational benefits of the tool are negated and may even lead to negative consequences.

Supplementary Notes on Large Effect Sizes

In this study, the effect sizes for phonetic recognition and the external load dimension were relatively large. Taking into account the characteristics of the pilot study and the intervention design, this can be explained from three perspectives. On the one hand, the traditional pure audio-shadowing model used in the control group, which lacks feedback and interaction, is inherently a less efficient baseline condition; the two-way interaction and immediate feedback in the experimental group formed a stark contrast, thereby amplifying the differences between the groups. On the other hand, as this study is a small-sample pilot study, the small sample size may have overestimated the effect size for the actual population to some extent; the effect size may decrease when large-scale empirical studies are conducted in the future. Furthermore, the control group in this study employed a pure audio-shadowing approach without any immediate feedback; its teaching efficiency is inherently lower than that of the blended model typically found in conventional classrooms, which involves teacher-led reading combined with individualised correction. This may have contributed to widening the inter-group differences to some extent. Should future studies adopt a control group that more closely resembles typical classroom practice, the effect size is expected to diminish. Consequently, the conclusions of this study should be regarded as preliminary explorations, and caution is advised when generalising them to a broader context.

5.2. Dialogue with Existing Research

The findings of this study align with the core tenets of existing research on multimodal learning, namely that well-designed multimodal approaches can enhance learning outcomes and the learning experience under specific instructional comparison conditions. At the same time, this study makes three contributions to the existing literature. Firstly, it refines the functional boundaries of the auditory modality. Previous studies have generally discussed the positive effects of multimodal literacy in broad terms. Through an analysis by question type, this study has clarified the core value of voice interaction in the phonetic dimension and its limitations in the graphemic dimension, providing a basis for the precise design of multimodal literacy programmes. Otsu and Izumi [3] research on the promotion of graphemic recall through voice assistance also indirectly corroborates the processing advantages of the auditory modality, though the two studies focus on different aspects of the process. Secondly, it revises the linear assumption regarding cognitive load. Existing research often assumes that reducing load automatically enhances effectiveness; however, by differentiating cognitive load measurements, this study found that a decrease in external load does not necessarily lead to a corresponding increase in internal load. This reveals the conditional nature of cognitive resource allocation in younger children and enriches the empirical evidence for the application of cognitive load theory in this age group. Altmeyer *et al.*'s [8] study on cognitive load measurement in children primarily validated the discriminant validity of distinguishing between internal and external load, whilst maintaining a cautious stance regarding the reliability of measuring related load in younger age groups; the findings of this study align with this position. Thirdly, it supplements the perspective on group differences. While existing research often reports average effects, this study, through academic stratification analysis, reveals the differential impact of voice interaction on students with varying academic backgrounds, reminding practitioners to be wary of the risk that technology may widen the gap.

5.3. Implications for Teaching

Based on the above findings, this study offers three targeted practical implications for literacy teaching in the early years of primary school.

First, adhere to the principle of modal complementarity and clarify the tool's role. Voice interaction tools should be positioned as aids for phonetic correction and must not be relied upon alone to undertake the full task of literacy instruction. In the classroom, they must be combined with visual and kinesthetic activities such as the decomposition of Chinese character components and writing practice to form a complete, multi-modal closed loop encompassing phonetic correction, writing formation, and the derivation and extension of meanings. This approach prevents weak character form processing resulting from over-reliance on the auditory modality.

Second, provide in-depth guided tasks to facilitate cognitive transformation. When using speech tools, teachers must not stop at ensuring correct repetition; they must simultaneously design thought-provoking questions and extension tasks—such as forming a word with the character or constructing a sentence using that word. By providing scaffolding, they guide students to channel their cognitive resources into understanding character meanings and applying them in new contexts, ensuring that the reduction in workload genuinely serves to enhance learning efficiency.

Thirdly, implement tiered feedback strategies to safeguard motivation in lower-achieving groups. Adopt differentiated assessment strategies for students of varying academic levels: for high-achieving groups, raise standards and emphasise extension activities; for students with weaker foundations, lower the initial difficulty, reduce the frequency of consecutive corrections, and increase the proportion of encouraging feedback, prioritising the protection of learning confidence before gradually raising accuracy requirements. This tiered instructional strategy is also consistent with the findings of differentiated instruction research in primary school phonetic teaching [11]. At the tool design level, difficulty levels and error-tolerance mechanisms should also be incorporated to avoid a one-size-fits-all assessment standard.

6. Conclusions and Outlook

6.1. Research Findings

This pilot study, conducted with Year 2 primary school pupils, examined the impact of voice-interactive multimodal teaching on Chinese character literacy outcomes, cognitive load and learning interest. The main conclusions are as follows:

1) Voice-interactive multimodal teaching significantly improves overall literacy scores, though there are differences across question types: it greatly enhances phonetic recognition, has limited benefits for character shape discrimination, and provides a moderate improvement in character-meaning matching.

2) This teaching model significantly reduces students' external cognitive load, but the associated cognitive load does not increase correspondingly; reducing cognitive load does not automatically lead to deep learning. This result may also be influenced by the validity of self-reported measures of associated cognitive load.

3) This teaching model generally enhances learning interest, but group differences exist: interest increased significantly in the high-achieving group, whilst the low-achieving group showed no significant gain; frequent error correction may lead to feelings of frustration.

Overall, the findings of this pilot study indicate that lightweight voice-interactive multimodal tools can produce positive immediate effects on phonics learning outcomes, extraneous cognitive load reduction, and learning interest enhancement when compared with a pure audio-repetition teaching mode. However, the functional boundaries of such tools must be clearly defined, and they should be paired with targeted instructional design and differentiated support strategies to achieve comprehensive literacy teaching benefits in real classroom contexts.

6.2. Limitations of the Study

As a pilot study, this research has five limitations: Firstly, sample limitations. The sample size was small and drawn from a single school, limiting the generalisability of the conclusions. Secondly, limitations regarding the intervention duration. The intervention lasted only two weeks, allowing observation of short-term effects only and failing to verify long-term retention rates. Third, the intervention modality was limited. The study examined the effects of the auditory modality alone, without exploring the combined effects of other modalities such as visual or kinesthetic input. Fourth, limitations regarding measurement tools. The cognitive load dimension was measured via self-report by the children; however, younger pupils have limited ability to understand and distinguish abstract concepts related to active construction efforts. Consequently, the validity of this dimension is weaker than that of the other two dimensions, which may affect the accuracy of the results. Furthermore, qualitative interviews were conducted only within the experimental group; as no experiential data from the control group was included, no qualitative comparisons between groups could be made. Future research could incorporate interviews with the control group to enrich the evidence triangle. Future research should incorporate objective behavioural indicators, such as reaction times and eye-tracking, to conduct a more precise, multi-method assessment of cognitive load. Fifth, limitations of the stratified design. No stepwise intervention programme was designed for the low-performing group, making it impossible to verify the actual effectiveness of the stratified feedback strategy.

Future research could be expanded in four directions: Firstly, increasing the sample size and the range of schools to enhance the generalisability of the findings; Secondly, exploring multimodal combinations that integrate speech, gestures and writing to identify the optimal model for the synergistic processing of form, sound and meaning; Thirdly, designing stratified intervention programmes to verify the motivational protection effects of stepwise feedback strategies on students in the low-performing group; Fourth, integrating multimodal learning analysis technologies to assess cognitive load and learning status more accurately using objective data such as speech features and operational behaviours.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Olmanson, J. and Liu, X. (2017) The Challenge of Chinese Character Acquisition: Leveraging Multimodality in Overcoming a Centuries-Old Problem. *Emerging Learning Design Journal*, **4**, 1-9.

[2] Mayer, R.E. (2005) The Cambridge Handbook of Multimedia Learning. Cambridge University Press. https://doi.org/10.1017/CBO9780511816819

[3] Otsu, K. and Izumi, T (2022) Interactive Handwriting Device for Enhancing Active Recollection of Character Forms by Voice Assistance for Chinese Character Learning. *Extended Abstracts of the* 2023 *CHI Conference on Human Factors in Computing Systems*, Hamburg, 23-28 April 2023, 1-7. https://doi.org/10.1145/3544549.3585814

[4] Yu, J., Song, J. and Lu, Y. (2025) Harnessing Generative Artificial Intelligence to Con-

struct Multimodal Resources for Chinese Character Learning. *Systems*, **13**, Article 692. https://doi.org/10.3390/systems13080692

[5] Sweller, J. (1988) Cognitive Load during Problem Solving: Effects on Learning. *Cognitive Science*, **12**, 257-285. https://doi.org/10.1207/s15516709cog1202_4

[6] Paas, F., Renkl, A. and Sweller, J. (2003) Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*, **38**, 1-4. https://doi.org/10.1207/s15326985ep3801_1

[7] Klepsch, M. and Seufert, T. (2020) Understanding Instructional Design Effects by Differentiated Measurement of Intrinsic, Extraneous, and Germane Cognitive Load. *Instructional Science*, **48**, 45-77. https://doi.org/10.1007/s11251-020-09502-9

[8] Altmeyer, K., Barz, M., Kapp, S. and Lauer, L. (2023) Digital Ink and Differentiated Subjective Ratings for Cognitive Load Measurement in Middle Childhood. *British Journal of Educational Psychology*, **93**, 368-385. https://doi.org/10.1111/bjep.12595

[9] DeLeeuw, K.E. and Mayer, R.E. (2008) A Comparison of Three Measures of Cognitive Load: Evidence for Separable Measures of Intrinsic, Extraneous, and Germane Load. *Journal of Educational Psychology*, **100**, 223-234. https://doi.org/10.1037/0022-0663.100.1.223

[10] Braun, V. and Clarke, V. (2006) Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, **3**, 77-101. https://doi.org/10.1191/1478088706qp063oa

[11] Chen, S.L. and Zeng, S.J. (2024) Effects of Differentiated Instruction on First-Grade Phonetic Teaching. *Journal of Textbook Research*, **17**, 125-158. (In Chinese)