The effects of different approaches to teaching Chinese orthography on character acquisition and reading performance in learners of Chinese as a foreign language: A systematic review

Sonia Qin

MSc in Applied Linguistics for Language Teaching, 2025

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:	Sonia Qin
Date:	30/09/2025

Acknowledgements

This dissertation would not have been possible without the invaluable support and guidance of my supervisor, Hamish Chalmers. His depth of knowledge and encouragement gave me confidence in the value of my work. He has inspired me to approach research with integrity and has helped shape this dissertation into something far better than I could have achieved alone. I learned a lot. Thank you!

A big shoutout to my ALLT cohort and peers. Our Sunday meetings made me feel part of a truly special community. Your kindness and generosity in sharing ideas and feedback were always constructive, and it was clear that everyone genuinely wanted to see each other succeed. I loved meeting everyone during induction week, and it was so wonderful to see some of you again in person in Australia! In addition, thank you to my family for their love and support throughout this journey. I am deeply grateful for their patience, understanding, and faith in me.

Finally, thank you, Michael. Your belief in me and your constant encouragement gave me strength during the most difficult moments. Even when I doubted myself, you helped me move forward and see the bigger picture. I could not have done this without you.

Abstract

There is growing evidence investigating the effects of different instructional methods on character acquisition and reading performance in contexts where students are learning Chinese as a foreign language (CFL). However, efforts to present a comprehensive and transparent synthesis of the literature have been limited, with previous systematic reviews suffering from incomplete reporting and the absence of risk of bias assessments. This makes it difficult to guide effective research and leaves educators unable to draw meaningful conclusions to inform sound teaching practice. Therefore, a critical evaluation of the current state of knowledge is needed, especially given the global expansion of Mandarin as the predominant variety taught in educational settings, and the inherent challenges associated with learning Chinese orthography.

A systematic search process identified 30 eligible studies, published between 1990-2024. Findings revealed five broad categories of pedagogical approaches: computer-assisted language learning, conventional (or traditional) techniques, holistic approaches (combining reading, writing, listening, speaking), timing of character teaching, and various presentation methods when introducing unfamiliar characters. Compared to conventional strategies like rote memorisation, studies generally favoured technology-enhanced or multimodal methods, holistic approaches, and different ways of presenting target items. Research on the timing of orthographic instruction was mixed, with some evidence promoting immediate teaching of characters over a pinyin-first approach. However, trustworthiness ratings indicated that the overall strength of the literature was generally weak, with 23 studies classified as having high risk of bias. The actual or potential effects of each approach, considered alongside its study quality, are crucial factors for CFL practitioners when deciding whether to maintain existing teaching strategies in the classroom or adopt those found in this review. Some instructional methods may be successful in certain learning contexts, but the overall low methodological quality of the available evidence makes it problematic to draw clear causal inferences and offer confident pedagogical recommendations. Future research must employ more robust and transparent study designs to support the development of evidence-informed CFL teaching.

List of Figures

List of Tables

List of Abbreviations

CALL	Computer-assisted language learning
CFL	Chinese as a foreign language
CLIL	Content and Language Integrated Learning
HL	Heritage learner
HSK	Hanyu Shuiping Kaoshi
L1/L2	First/second language
FL	Foreign language
IDESR	International Database of Education Systematic Reviews
OVAL	Observe, visualise, articulate, listen
RCT	Randomised controlled trials
RoB	Risk of bias
SLA	Second language acquisition
UK	United Kingdom

Table of Contents

Chapter 1: Introduction

1.1 Aim and rationale

With China's growing prominence on the global stage, the teaching and learning of Chinese as a foreign language (CFL) at both school- and university-levels have gained popularity among students from both Anglophone and non-Anglophone regions (British Council, 2023; Gil, 2024). Despite rising enthusiasm, students often encounter significant challenges when learning CFL. This is well-documented in Anglophone contexts, where poor outcomes and discouraging experiences have partly contributed to high attrition rates when students decide whether to pursue the language at more advanced levels (Yang, 2022; Yue, 2017). Potential factors influencing these sentiments and underperformance include the tonal nature of Mandarin (Kan et al., 2018) and limited curriculum time allocated to CFL learning in formal settings (Orton, 2016). Character writing is also widely acknowledged as one of the greatest obstacles for many learners, especially those from alphabetic language backgrounds due to its orthographic distance (Duff et al., 2013).

However, systematic reviews on effective strategies for teaching Chinese orthography among CFL learners remain scarce. Prior reviews (e.g. Li, 2020; Zhang, 2024) lack rigour, largely due to incomplete reporting and the absence of risk of bias (RoB) assessments, minimising their practical value for educators seeking to make evidence-informed pedagogical decisions. There is also limited clarity regarding which teaching methods have been evaluated for effectiveness across various proficiency levels, with most research focusing on beginners. The present study employs a systematic review approach to gather and assess all available evidence on CFL teaching, examining both the nature and extent of this research, and the effectiveness of different approaches on character acquisition and reading performance. These outcomes have been chosen because anecdotal data suggest that students frequently struggle with learning and retaining characters (Walker & Poole, 2016; Yang, 2022). Additionally, the absence of an alphabet in Chinese and its morpho-syllabic writing script requires learners to integrate the character's sound, meaning, and visual form to successfully recognise and comprehend its meaning (Tong & Yip, 2015). As Chinese contains many spoken varieties (e.g. Cantonese), this review focuses specifically on Mandarin pronunciation, which is most commonly taught in schools (Goh, 2017). Examples of the writing script in the following chapters are presented in simplified characters.

1.2 Outline

Chapter 2 explores in greater depth the growing interest in CFL education, the challenges with learning Chinese and its orthography, and compares teaching methods between native Mandarin speakers and CFL students. It then discusses the relationship between CFL and second language acquisition (SLA) research, followed by an analysis of previous systematic reviews. The chapter concludes by introducing the study's research questions. Chapter 3 outlines the methodology and Chapter 4 presents findings from the search process. Chapter 5 narratively synthesises the evidence, while Chapter 6 discusses the review's limitations and implications for research and teaching.

Chapter 2: Background and rationale

2.1 The global spread of Mandarin

As the world's second most-spoken language, Mandarin has accumulated over one billion native speakers (Goh, 2017). It is the official language of China and Taiwan, and is widely used as the lingua franca among the Chinese diaspora (Goh, 2017). Several factors have contributed to the expansion and increasing prominence of Mandarin as a world language, including the effects of globalisation and international migration, as well as China's political and economic dominance in recent decades (Erbaugh, 2022). In the United Kingdom (UK), Mandarin has been identified as the fourth most important language for the country's future over the next 20 years (Tinsley & Board, 2013). Efforts to increase the number of non-native speakers in public schools have included the Mandarin Excellence Program, which supports British secondary students in developing proficiency from an early stage (Nicoletti & Culligan, 2022). Recent data indicate that 6% of public secondary schools in England now teach Mandarin as a Key Stage 3 curriculum subject, with approximately 1% of Year 11 students achieving a qualification – representing modest numbers (British Council, 2023). Similarly, the Flagship Program in North America offers opportunities for university students to enrol in short-term intensive courses focused on achieving high levels of fluency, which supposedly cannot be reached through regular foreign language (FL) classrooms (Everson, 2011). In Australia, Mandarin has become an integral component of the national curriculum in fostering 'Asia literacy'. This policy aims to develop young people's knowledge and appreciation of Asian countries, equipping them with the skills to live and work in the region (Asia Society Australia, 2022). At the same time, Mandarin has been offered as a curriculum subject in Ireland's Leaving Certificate since 2017 (Osborne et al., 2022).

The rising number of non-native learners has also extended to non-Anglophone regions, with Uganda recently introducing Mandarin as a compulsory subject in its high-school syllabus (Xu, 2023). Kenya and South Africa have taken similar steps by offering the language as an elective in public schools, thereby illustrating its growing importance in African society (Gil, 2024). This enthusiasm can be attributed to the perceived status of China as a "global superpower" (Xu, 2024, p. 4) and the country's substantial investment in the continent, including cheaper tuition fees for African students attending Chinese universities (Gil, 2017).

The Chinese government has continued to promote CFL education through its expansion of Confucius Institutes and Confucius Classrooms. Confucius Institutes are established through partnerships between Chinese and foreign universities to facilitate cultural exchange and CFL learning at the tertiary level (Hartig, 2012). Meanwhile, Confucius Classrooms concentrate on primary and secondary education in public and local community schools (Starr, 2009). By the end of 2019, there were around 550 Confucius Institutes and 1150 Confucius Classrooms in 162 countries (Gil, 2024). It is clear that significant resources continue to be invested in Mandarin teaching and learning in both Anglophone and non-Anglophone regions. Therefore, it is important to evaluate the most effective methods for teaching Mandarin, given increasing efforts to improve non-native speakers' proficiency.

2.2 Chinese orthography

2.2.1 Terminology

Chinese comprises several varieties that are mutually unintelligible, including Mandarin, Cantonese, Min, Wu, Xiang, Gao, and Hakka (Chen, 1992). The main difference between Mandarin and other varieties lies in its usage: Mandarin (or Putonghua, meaning 'common speech') is based on the pronunciation adopted in Beijing. It is the standard national variety endorsed by the Chinese government, despite nearly 30% of the population being unable to speak it (Luo, 2014). Conversely, varieties like Cantonese are largely used in informal and intimate settings, such as daily conversations (Snow, 2013). While all Chinese varieties share the same orthography, their spoken forms differ significantly in lexis and tonal systems, with the number of tones ranging from three to ten (Xing, 2006). In this review, the term 'Chinese' pertains to its orthography/writing system and is used more broadly when discussing CFL education. Meanwhile, 'Mandarin' refers to the standardised spoken form most commonly taught in educational institutions globally. This study uses Hanyu pinyin – the romanisation system for transcribing Mandarin.

2.2.2 The writing system

Chinese is often considered the most linguistically distant from alphabetic languages due to its unusual writing system (Perfetti & Liu, 2005). Instead of letters representing sounds, the Chinese language uses characters to convey meaning (Lü & Zhang, 1999). The composition of each character contains various strokes and radicals, arranged within a square structure (Shen, 2005). Strokes are the individual lines that shape each character. They are typically written systematically from top to bottom and left to right (Chen, 1992). Radicals are the basic components of a character and can provide clues about its pronunciation or meaning: phonetic or semantic radicals respectively (Chen, 1992). A single radical can represent a whole character (called an integral character), whereas combining two or more radicals produces a compound character (see Figure 2.1). Boltz (1994) categorises the structural framework of characters into the following: pictograms, ideographs, ideographic compounds, phonetic loan words, phono-semantic compounds, and derivative cognates (see Table 2.1).

Moreover, Chinese includes two distinct writing scripts: simplified and traditional. Simplified characters were introduced by the Chinese government to reduce native speakers' illiteracy and enhance overall education (Snow, 2013). While China has largely adopted the use of simplified characters, traditional characters remain the official script in Hong Kong, Taiwan, and Macau (Goh, 2017). CFL students usually learn simplified characters, as they are less visually complex and involve fewer stroke sequences to produce and memorise (Ésik, 2020). This review uses simplified characters for discussions on Chinese orthography. However, when referring to experimental conditions in empirical research, the type of script used will be clearly indicated.

Fig 2.1

Layout of a compound character (adapted from Ye & McBride, 2022)

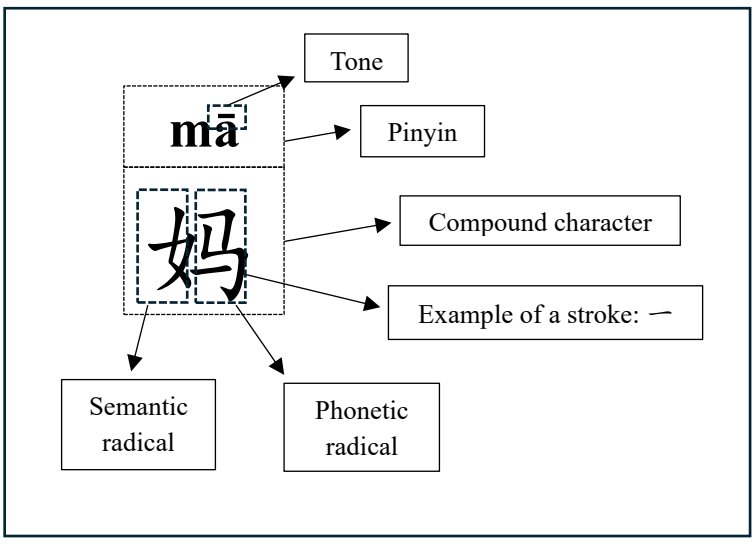


Table 2.1

Composition of Chinese characters (adapted from Boltz, 1994)

Type	Explanation	Example
Pictograms	A visual representation of the word	 mù = 木 (wood)
Ideographs	Characters that visually represent abstract concepts without a physical form	Numbers one, two, and three are depicted by the same number of strokes: yī èr sān 一，二，三
Ideographic compounds	Combines two or more pictograms to create a new meaning/character	 lín = 林 (forest)
Phono-semantic compounds	Contains both phonetic and semantic radicals in one character	Mother = 女 (nǚ) + 马 (mǎ) = 妈 (mā); 女: Semantic radical; 马: Phonetic radical
Phonetic loan words	Words that sound like their foreign counterparts. They are borrowed from other languages (mainly English) and usually include modern non-Chinese names/countries	jiā ná dà 加拿大 (Canada); bǐ ěr gài cí 比尔盖茨 (Bill Gates)
Derivative cognates	Characters that share the same etymological root, but differ in pronunciation and meaning	lǎo 老 (old) vs. kǎo 考 (to test)

2.3 Why is Chinese difficult to learn?

2.3.1 Linguistic and psychological challenges

It is widely assumed that learning CFL is especially challenging for students from alphabetic language backgrounds (Hu, 2010). A significant source of this difficulty lies within the ability to read and write its characters. Indeed, one must be able to recognise over 3000 characters to achieve ‘basic’ proficiency (Norman, 1988). For example, a beginners course at Oxford University introduces students to approximately 130 characters per term, with two-hour tutorials per week over 10 weeks (Department of Continuing Education, n.d.). These courses are primarily designed for non-native adult learners with limited prior knowledge. If a student underwent four terms of university classes studying around 500 characters per year, it would take them nearly six years to reach the target of 3000 characters, provided all content taught

is retained and there is minimal out-of-school exposure. The cognitive effort placed on one's memory are similarly noted by British CFL secondary learners, who describe the task of recalling "every single character" as nearly "impossible" (Yang, 2022, p. 8). Even when compared to a first-language (L1) English-speaking student, native Chinese children often spend an additional two years developing fundamental literacy skills (Lü & Zhang, 1999).

CFL learners also study the pinyin romanisation system to effectively understand and speak the language (Kan et al., 2018). Each character represents one syllable in pinyin transcription. These syllables are typically shorter than English words, ranging from one to five letters and containing a limited set of sounds (Orton, 2016). This results in many homophones, making it more difficult for learners to disambiguate the correct meaning of words or characters when listening and speaking (Taylor & Taylor, 2014). Furthermore, pinyin spelling differs greatly from English, with almost 24% of pinyin words beginning with x, y, or z, compared to only 0.6% of English words starting with the same letters (Orton, 2016). CFL learners must become accustomed to writing pinyin with diacritical marks above the vowels of each syllable to support pronunciation (Everson, 1994). These marks (or tones) are part of the entire syllable rather than an innate feature of the vowel itself (Lin, 2007; see Table 2.2).

Table 2.2

Mandarin tones (adapted from Lin, 2007)

Pinyin	Character	Tone number	Meaning	Pitch sound
mā	妈	1	Mother	Level
má	麻	2	Hemp	Rising
mǎ	马	3	Horse	Dipping
mà	骂	4	To scold	Falling
ma	吗	5	Question word	Neutral

Another barrier in its structural complexity is that phonetic and semantic radicals do not always match the character's intended pronunciation or meaning (Honorof & Feldman, 2006). Phonetic radicals are approximately 38% reliable in predicting the character's sound, excluding tone variations (Perfetti et al., 1992). However, this falls to 26% when the tone of the phonetic radical corresponds to the tone of the entire character (Wong, 2013). For example, the phonetic radical in 饿 (hungry) is 我 /wǒ/ (I, me), but the correct pronunciation is /è/. The semantic radical on the left indicates a connection to food, but it cannot function alone as an integral character. As phonemic cues in written Chinese are not always reliable, CFL learners who heavily depend on such strategies to decode words/phrases may find this aspect of reading more challenging (Mori, 1998; Shen & Ke, 2007).

Besides lacking an alphabet, CFL students face additional linguistic hurdles due to the language's extensive vocabulary and four-character expressions, many of which have no direct English equivalents and typically hold cultural/historical significance (Chen, 1992). Due to the heavy reliance on context for syntactic information in Chinese sentences, even advanced CFL learners struggle with reading texts that are not specifically tailored to their proficiency (Huang, 2022). Therefore, learners must acquire a diverse range of vocabulary to convey and understand even simple ideas (Duff & Li, 2004). Most characters are also lexical morphemes, meaning they can function both as standalone words or combine to form multi-character words that represent only one English word (Lü & Zhang, 1999). For example, the English word 'friend' is represented by two characters: 朋友 /péng yǒu/. This difficulty is worsened by the uniform spacing between adjacent characters in Chinese writing. That is, because there are no spaces between characters, it is easy to mistake a character for an entire word, when in reality, it is simply part of a multi-character word. Thus, orthographic learning is an intensive process that demands substantial time, memory, and cognitive effort.

2.3.2 Systemic challenges

While learning Chinese presents both linguistic and psychological obstacles, other external factors also contribute to such challenges, including insufficient curriculum time in schools and poorly-designed teacher education programs (Hao & Li, 2024). The American Foreign Service Institute estimates that it takes around 2200 class hours for CFL pupils to achieve working-level proficiency in Mandarin, compared to only 750 hours for European languages like Spanish (US Department of State, n.d.). It is understandable that institutions are reluctant

to devote such extensive curriculum time to teaching and learning one foreign language. For example, language studies in Sydney, Australia are only mandatory in Years 7-8 (aged 12-13), amounting to roughly 100 contact hours (NSW Education Standards Authority, n.d.). Attrition rates in Mandarin classes are notoriously high in Australian secondary schools, with only 2.5% of CFL students continuing through to Year 12 (Orton, 2016). This situation can be attributed to limited contact hours, students' slow progress in acquisition, and the fact that learners with minimal CFL knowledge must compete against heritage students at the Higher School Certificate (Orton, 2011). In Australia and America, many institutions do not differentiate between CFL and Chinese Heritage Programs due to low enrolments (Orton, 2011; Yue, 2017). Heritage learners (HLs) are individuals who have been exposed to a language from childhood, typically through family, but are not considered native speakers due to a shift in the majority language used before adulthood (Valdés, 2001). These learners usually develop some degree of proficiency in listening and speaking, but often lack reading and writing skills (Murphy, 2014). Thus, children learning Chinese for the first time often feel demotivated as they are studying alongside HLs, whom they perceive as already fluent native speakers (Orton, 2011). While research on this phenomenon in non-Anglophone regions is scant, the situation appears more promising in the UK. It is reported that students in the Mandarin Excellence Program attend four hours of class per week as part of their school timetable, along with an additional four hours outside of class (British Council, 2024). Yet, students progressing to Mandarin A-levels are predominantly native speakers at independent schools (Nicoletti & Culligan, 2022).

Chinese teacher education programs have also faced criticism for failing to address the specific linguistic needs involved in CFL education (Zeichner & Liu, 2010). Although some universities worldwide now offer specialised credentialling courses and education degrees for teaching Chinese, many preservice teachers still follow generic language training programs (Everson & Xiao, 2009; Yue, 2017). These programs focus on developing general syllabus knowledge, often overlooking the pedagogical and linguistic skills required to effectively teach students from diverse learning backgrounds, including those with disabilities (Kwoh, 2007). This issue is especially prevalent in Australia (Orton, 2016), America (Yue, 2017), and the UK (Wang & Higgins, 2008), with Wang and Higgins identifying the dearth in suitable teaching materials as the main challenge faced by CFL practitioners in the UK. Orton (2016) echoes similar sentiments, stating that resources are woefully underdeveloped in Australia.

It is clear that characters are a major challenge for CFL teaching and learning. Nonetheless, they remain an essential component for students to practise and memorise, especially given their already limited exposure to the spoken form. Character acquisition ultimately involves mastering three key components: shape (orthography); sound (phonology); and meaning (semantics). Consequently, it is crucial to examine the extent to which different instructional methods influence CFL learners' reading performance and character acquisition. In doing so, this allows educators to combine their own expertise and experiences with robust evidence-based practices when making pedagogical decisions (Chalmers, 2016).

2.4 Teaching methods

2.4.1 Native Chinese students

According to Lam (2011), native speakers traditionally adopt two main approaches to teaching Chinese orthography: the character-centred (intensive) and meaning-centred (extensive) approach. The intensive approach focuses on teaching individual characters, supporting students to cultivate a large vocabulary before developing literacy skills. Popular methods aligned with this approach include teaching characters through rote memorisation and repetitive handwriting practice. However, this approach requires students to learn every single character before the reading process, which may restrict them to simpler or less challenging materials as the content is limited to the range of characters they have mastered (Liu & Lo Bianco, 2007). In contrast, the extensive approach focuses on teaching characters through meaningful texts, such as poems and short stories. In other words, students gradually learn characters through reading. This typically involves incorporating pinyin to support early reading development, rather than initially focusing on linguistic patterns within the characters themselves. However, since Chinese children already possess familiarity with the spoken language before learning to read and write, these approaches may arguably be less effective for CFL learners. The following section explores some methods for teaching characters to CFL students and discusses the theoretical rationales underpinning such approaches.

2.4.2 Timing of orthographic instruction

There has been considerable debate regarding the optimal time to introduce and teach characters in CFL classrooms (Everson, 1988; Walker & Poole, 2016). Some proponents (Packard, 1990; Swihart, 2004) argue that introducing characters should be delayed, promoting a 'pinyin-first approach' to build learners' confidence in aural-oral skills. This perspective is based on the speech primacy theory, which posits that a strong foundation in

speaking and listening is essential for literacy development (Ye, 2013). Swihart (2004) claims that CFL practitioners should prioritise teaching the sound system, vocabulary, and grammar through pinyin before reading and writing characters, though this risks diminishing students' motivation to learn the writing system and fostering an over-reliance on pinyin.

Packard (1990) found that delaying character instruction improved oral proficiency with no observed benefits in character production when compared to those who underwent immediate exposure. Participants were enrolled in a CFL beginners course at an American university and were divided into two groups: the 'delayed' group (n=12) received instruction from the fourth week of the 13-week semester, while the 'early' group (n=11) were exposed to characters from the first week. They were taught traditional characters by native Mandarin speakers. Assessment measures included pinyin transcriptions, English-to-Chinese translations, and a paired conversation where students conversed in Mandarin based on provided prompts. Both groups were only tested on the characters they had learnt, but the delayed group attended three additional one-hour lessons per semester. Data were collected twice throughout the year: T1 (end of first semester) and T2 (end of second semester).

The author reported no significant between-group differences in students' scores at T1, though the delayed condition performed significantly better in oral proficiency and pinyin transcriptions at T2. He suggests that delaying instruction facilitates the development of oral skills and phonetic discrimination, while early exposure offers no significant advantage on any measure. However, one might argue that the delayed group would naturally perform better since they received supplementary lessons devoted to character instruction. It is unclear whether participants' speaking gains are solely due to the delayed start or simply the result of additional exposure. McGinnis (1999) contends that early instruction is mostly implemented in actual CFL classrooms, stating that more curriculum time would be necessary to establish connections between its orthography and meaning. This calls into question the practicality of adopting a delayed curriculum, especially given the already limited time allocated to CFL instruction (see Section 2.3.2 for an overview of the broader systemic challenges facing CFL education).

2.4.3 Radical components and chunking

Research indicates that understanding radicals and their semantic/phonetic functions in character formation can enhance orthographic awareness among CFL adult learners (Shen, 2005; Taft & Chung, 1999). It is assumed that when students perceive characters as structured components, it is easier for them to visually encode and process the entire character (Chang et al., 2014). Rather than viewing characters as disorganised strokes, recognising meaningful units (i.e. radicals) supports learners to form larger, more manageable chunks of information (Ellis, 2003). This process, known as chunking, involves grouping smaller elements in memory into larger cohesive units to increase information capacity (Newell, 1990).

Xu et al. (2014) investigated whether grouping characters by common radicals influenced orthographic processing and character recognition across different proficiencies. The authors recruited 48 first-year beginners and 40 second-year intermediate-level students enrolled in CFL courses at an American university. Participants were divided into two conditions, but it is unknown how group allocation was determined: 24 beginners and 20 intermediate learners were assigned to the grouped condition (characters were taught using shared radicals), while the remaining were placed in the distributed condition (orthographic instruction was spread across eight radical groups). The intervention comprised four one-hour sessions conducted by participants' regular teachers. Forty radicals and 48 simplified characters derived from those radicals were taught, using identical vocabulary and reading texts across conditions. Data collection included pre-post- and two-week delayed post-tests, covering lexical decision and matching tasks, pinyin transcriptions, English-to-Chinese translations, and radical recognition and semantic awareness exercises. The latter required students to identify the radical's visual form and semantic function in unfamiliar characters.

The results demonstrated that beginners in the grouped condition scored significantly higher across all assessment measures. However, this advantage was temporary with no significant differences observed over time. Intermediate learners in the grouped condition only showed short-term non-significant improvement in radical awareness, pinyin, and translations, but no lasting benefits were observed on any measure. The authors offer three explanations for these findings: first, the teaching materials contained high-frequency radicals that intermediate learners may have previously encountered; second, these learners likely had greater exposure to radicals and CFL learning overall, having studied the language for two years; and third, the instructional duration was too short to produce significant effects. The authors conclude that

radical-grouping is successful for students with "no more than a few weeks of experience in Chinese orthography" (Xu et al., 2014, p. 788). They further argue that repeated exposure to radicals can enhance intermediate learners' ability to apply radical knowledge to novice characters, but this benefit is only short-term. However, all groups used the same reading texts during the intervention, raising concerns about the suitability of the teaching materials. The content may have arguably been too difficult for beginners or too simple for intermediate learners, potentially impacting the validity of results. It also suggests that differences in outcomes may not stem from the intervention but rather could be affected by the mismatch between CFL proficiency and text difficulty. Nonetheless, the findings indicate that radical-grouping can facilitate temporary gains in orthographic knowledge.

2.4.4 Handwriting practice

While often considered a traditional approach, handwriting practice remains a contentious topic in CFL pedagogy (Lü et al., 2019). Some critics (DeFrancis, 1984; Walker, 1989) contend that memorising how to write characters by hand is "extremely inefficient" (Allen, 2008, p. 237), especially in the early stages of CFL learning. This is because beginners lack the linguistic foundation needed for successful memorisation and reading, contact hours are severely limited, and prioritising handwriting offers fewer practical benefits for real-world use in the current digital age (Allen, 2008). Indeed, some studies have illustrated that typing through pinyin input, as an alternative to handwriting, is equally effective in improving text production length and word recognition (Chen et al., 2017; Zhang, 2021). However, other scholars (McBride, 2015; Packard et al., 2006) argue that handwriting is an essential aspect of the character-learning process. Research involving native Chinese children and CFL adult pupils indicate that repetitive handwriting practice facilitates orthography-semantics mapping (Cao et al., 2013; Hsiung et al., 2017). The assumption is that handwriting allows learners to focus on the visual structure and stroke order/composition of characters, connecting sensory-motor information with mental representations (Tan et al., 2005).

Harvey and Brooks (2022) provide evidence for the benefits of handwriting, suggesting that primary-school learners achieve greater literacy outcomes. The authors recruited 70 Grade 4 students enrolled in a Chinese immersion public school in America. Participants' L1s were not disclosed, as these details were not provided by the school. All students had received three years of prior Chinese-language instruction and were considered beginner learners. Intact classes were assigned to conditions based on teachers' decisions: two groups engaged

in typing via text-messaging, while another two groups completed handwritten tasks. All groups attended twice-weekly sessions for eight weeks (15-20 minutes per lesson), taught by different instructors. In the typing condition, students used individual computers to discuss topics (e.g. family) based on the provided prompts, in simplified characters. Conversely, the handwriting group completed paper-based worksheets, such as counting strokes and writing exercises. Reading skills and character production were evaluated pre-post- intervention using Level Chinese (a standardised reading assessment) and free-writing activities, where students handwrote as much as possible.

Findings illustrated no significant differences in reading scores between conditions, though students engaged in handwriting still performed better than their peers. Free-writing results indicated that the typing group performed comparatively worse, producing fewer characters and demonstrating less lexical diversity. The authors conclude that learners' poor vocabulary and pinyin spelling hindered meaningful engagement in online conversations. Consequently, students often disregarded texting prompts and resorted to basic vocabulary to discuss unrelated topics. However, all assessments were completed in written format, which may have impacted the handwriting performance of students in the typing condition, as most of their time was spent on typing practice. Although findings were not statistically significant, the authors conclude that handwriting should remain a core component of CFL learning.

The above methods offer different approaches to teaching Chinese characters, exhibiting varying levels of success based on the assessment measures used. Evidence focusing on CFL education among novice students indicates that radical-grouping and handwriting can promote character recognition, production, and reading, while delaying orthographic instruction appears to facilitate oral proficiency. Despite cumulative evidence showing that radical instruction supports character learning (Wong, 2017; Zhang et al., 2016; Zhang & Ke, 2018), ongoing debates persist regarding the efficacy of immediate versus delayed character exposure (He, 2023; Walker & Poole, 2016), and the practical value of handwriting compared to typing (Zhang, 2021). There remains no consensus regarding the effects of different instructional approaches on CFL learners' character acquisition and reading performance, particularly across diverse proficiency levels, L1 backgrounds, and age groups.

It is also noteworthy that pupils in the preceding studies were self-selected into beginner programs at American institutions. This presents a possible bias, in that students who choose to study CFL at primary- or tertiary-level are generally more motivated and driven by internal and external factors, such as career opportunities in Asia, and a potential desire to integrate into Chinese society (Dos Santos, 2024; Mayumi & Zheng, 2023). This trend is particularly prevalent in Anglophone and European settings, where Mandarin is typically offered as an elective subject in schools, often attracting students with a personal interest in Chinese (Dos Santos, 2024; Xu & Moloney, 2019). Therefore, it is important to consider how such factors might affect the broader educational landscape of CFL, especially in efforts to make Mandarin more accessible to students from diverse L1 backgrounds and to promote proficiency beyond the beginner stage.

2.5 CFL learning and its relationship to SLA research

Research in FL teaching and SLA remains primarily focused on the English language (Ma et al., 2017). This raises questions as to whether CFL education is fundamentally 'unique' or if studying Chinese draws on the same psycholinguistic processes that underpin SLA more broadly (Han, 2016). In the context of this review, it is necessary to consider the extent to which SLA research/theories – largely developed through evidence on teaching English as a FL – are applicable to CFL pedagogy. For example, Chan et al. (2022) suggest that the 'dual route model' of word recognition, which plays a central role in explaining English reading processes (see Coltheart, 2005), has limited relevance for reading Chinese characters. This is because, for students who lack oral proficiency, the phonological route is far less accessible, as characters cannot simply be 'sounded out' to infer meaning. Thus, SLA theories may be insufficient for describing the processes involved in learning Chinese, and alternative or adapted frameworks may be necessary to account for such processes (Zhao, 2011).

2.6 Previous systematic reviews on teaching characters

Previous reviews on CFL instruction have varied widely in both their analytical approaches and research questions (Ma et al., 2017; Wang, 2025; Zhang, 2023). Li (2020) conducted a systematic review of articles published between 2005-2019, examining methodological trends and research themes in character teaching and learning both within and outside China. The author excluded papers from Hong Kong, Taiwan, and Macau, citing "sociopolitical and historical differences" in pedagogy (Li, 2020, p. 42). The resulting synthesis of 214 papers revealed that different aspects of character teaching and learning have been explored using

diverse methodological approaches. The largest body of research in China (100 out of 142) adopted non-empirical designs (e.g. state-of-the-art reviews), while most studies conducted outside China (59 out of 72) employed experimental designs, typically using questionnaires, tests, and surveys for collecting data. The research distribution of studies indicated that pedagogical strategies were the primary focus in China (n=74), while computer-assisted language learning (CALL) attracted the most interest outside China (n=28). Based on these findings, the author recommends conducting more empirical research to establish stronger connections between theory and practice. He also encourages more qualitative studies outside China to examine how personal factors like motivation may impact the acquisition process. Although Li (2020) provides valuable insights into the research trends in CFL pedagogy, especially within China, the study's replicability is weakened by the lack of transparency in the eligibility criteria. The absence of RoB assessments also makes it challenging to evaluate the overall strength and reliability of the evidence.

More recently, Zhang (2024) systematically reviewed 22 studies between 1952-2023, investigating research trends and effects of different methods for teaching characters on L1 English learners' writing outcomes. The synthesis of the included studies showed that various strategies have been employed to support students in learning characters, such as handwriting practice, teaching radicals in chunks, various ways of presenting character information, the timing of character instruction, and haptic teaching through body movements. The majority of research was conducted in university settings (17 out of 22), predominantly involving adult beginners in their first year of tertiary education. In contrast, significantly fewer studies were administered in middle- (n=2) and high-schools (n=3). Research was also distributed across three geographic regions: America (n=14); Australia (n=5); and Ireland (n=3), with more than half implementing experimental designs (n=12), seven using mixed-methods, and three being exploratory. The outcomes measured across studies were thematically categorised into character recognition, production, radical awareness, and pinyin recall. Zhang (2024, p. 22) concludes that several methods are effective in improving students' writing outcomes, stating that there is no "single best method" or "right way" to teach characters. However, this claim is weakened by the lack of clarity regarding which 'several methods' the author is explicitly referring to, as well as the failure to acknowledge that many of the included studies did not use a comparator group for evaluating effectiveness. Furthermore, the author did not assess the trustworthiness of the evidence or disclose the search terms used when looking for papers. The scope of the review is limited by its focus on a single population (L1 English students),

providing a monolingual perspective that excludes other groups of learners. Such incomplete reporting and selection bias hinders an informed evaluation of the overall strength of the findings.

The preceding reviews lack sufficient detail to draw strong conclusions about CFL pedagogy. It is further unclear which variety and writing script were taught in the included studies, as broad terms such as ‘Chinese language’ or ‘Chinese’ were often used to describe the target language – possibly reflecting inaccurate reporting by the original authors. The absence of quality appraisals makes it difficult to ascertain which methods are effective in improving linguistic outcomes at different proficiency levels. Hence, a more comprehensive, replicable, and transparent review of the current literature on CFL instruction is needed. Identifying successful strategies to enhance character acquisition and reading skills among CFL learners is valuable for educators, as it can inform future-related investigations and strengthen the link between research and practice. Without an unbiased evaluation of the existing evidence, the effectiveness of CFL teaching remains unclear, and teachers risk basing their pedagogical decisions on experience and intuition alone.

2.7 Research questions

(1) What instructional methods or pedagogical approaches to teaching Chinese orthography have been evaluated for effectiveness among school- and university-aged CFL learners?

    (a) What are the effects of these methods on character acquisition and reading performance?

(2) What methods or approaches to teaching Chinese orthography are most effective for CFL learners at different proficiency levels?

Chapter 3: Methodology

The purpose of a systematic review is to responsibly and transparently identify, gather, and appraise the totality of research evidence on a particular topic (Macaro, 2019). This enables stakeholders to access relevant findings in condensed formats, supporting both policy implementation and the development of effective pedagogical strategies (Gough et al., 2017). Systematic reviews differ from traditional literature reviews: the latter can be influenced by authors' biases, both in the selection of studies and how those studies are presented/critiqued (Newman & Gough, 2020). Authors' decisions in choosing certain studies for analysis can result in "selective [and] opinionated" reporting, as well as inconsistent representations of research (Oakley, 2007, p. 96). That is, literature reviews are often (1) selective, typically including studies that authors are familiar with, and (2) opinionated, reflecting personal views/biases rather than being objective (Oakley, 2007). Therefore, the value of systematic reviews derive from their implementation of clearly defined protocols, which maximises the chance that the review is an unbiased account of the current evidence, allowing for future updates and replication (Petticrew & Roberts, 2006). The chapter begins by outlining the review's eligibility criteria, information sources, search strategy, study selection and data collection procedures, and data extraction methods. It then describes approaches used for synthesising evidence and assessing RoB. This study was registered with the International Database of Education Systematic Reviews (IDESR, n.d.) (see Appendix A for protocol document).

3.1 Eligibility criteria

Table 3.1 specifies the inclusion and exclusion criteria for this review.

Table 3.1

Eligibility criteria

Item	Inclusion criterion	Rationale
Bibliographic information	Include 1: Studies with a complete reference. Exclude 1: Studies without a complete reference.	Complete bibliographic information is necessary for retrieval of full reports.
Date of publication	Include 2: No restrictions on publication date.	Collecting all relevant data, regardless of publication date.
Participants	Include 3: Studies focused on typically developing CFL learners, including those	This review evaluates the effectiveness of

	without explicit mention of special needs if typical development can be reasonably assumed. Exclude 3: Studies focused exclusively on non-typically developing learners (e.g. those with developmental language disorders or fine motor issues).	teaching methods in typically-developing students. Findings on non-typically developing learners may lack applicability to a broader population.
	Include 4: Studies in primary, secondary, and university contexts. Exclude 4: Studies conducted in informal settings (e.g. parents teaching at home) and learners not enrolled in formal language courses (e.g. learning for pleasure).	This review focuses on CFL learners' linguistic outcomes in formal settings only. Studies in informal settings or with learners not enrolled in language programs will be excluded.
	Include 5: Studies with CFL learners, including those without explicit mention of the variety, if reasonable to assume that pupils are studying Chinese orthography (simplified or traditional) with Mandarin pronunciation and pinyin. Studies that compare non-native and heritage students will be included, though the focus will be on non-native speakers' outcomes. Exclude 5: Studies exclusively with heritage learners, students in Chinese-speaking regions (e.g. Singapore, Hong Kong, Taiwan) or non-Mandarin varieties (e.g. Cantonese).	This review concentrates on Chinese orthography, using Hanyu pinyin and Mandarin pronunciation. Studies exclusively with heritage learners, students in Chinese-speaking regions, or non-Mandarin varieties will be excluded due to significant differences in prior experiences, lexis, and tonal systems.
Intervention	Include 6: Studies where learners receive some form of instruction, including studies where part of the teaching might involve independent work (e.g. homework). Exclude 6: Studies without an instructional period or focus on teaching methods (e.g. learner strategies).	This review aims to evaluate the effects of teaching methods to improve and inform CFL practice. Thus, the intervention must include an instructional component.
Outcomes	Include 7: Quantitative measures of reading performance and character acquisition (e.g. character/word recognition, orthographic knowledge, pinyin transcriptions, reading fluency and comprehension). Exclude 7: Only narrative analyses are offered without quantitative assessment measures on reading outcomes, or the focus is on non-linguistic aspects (e.g. motivation, attitudes).	While qualitative data can provide rich insights into whether a particular teaching method is well-received, the study focuses on identifying which teaching methods facilitate better linguistic outcomes.

Publication status	Include 8: All publication types (e.g. theses, dissertations, journal articles, conferences).	The current study includes all publication types to reduce potential publication bias.
Study design	Include 9: Experimental studies identifying causal relationships (e.g. randomised controlled design, quasi-experimental design, matched-pair design, regression discontinuity design). Exclude 9: Studies that do not explore causality or which are not primary research (e.g. observational studies, ethnographies, systematic reviews, qualitative research, state-of-the-art reviews, non-intervention studies).	This review aims to provide evidence on the effectiveness of different teaching approaches. Studies that explore causality are likely to retrieve relevant papers that evaluate the impact of an intervention.
Language of publication	Include 10: Publications in any language. Exclude 10: Do not exclude studies based on language of publication.	Exclusions of papers in any language may overlook an important body of literature that may help answer the research questions.

An important consideration was the language of publication, as constraining systematic reviews to English-only sources could have excluded literature relevant to the research questions (Moher et al., 2003). Recognising that research on CFL teaching and learning might identify studies in Chinese, titles/abstracts in non-English languages were included, provided they met the inclusion criteria. If eligible studies published in Chinese were identified, the reviewer who is proficient in the language conducted the screening and data extraction.

3.2 Information sources

To ensure a comprehensive review, the search for full reports covered databases from a range of academic disciplines (see Table 3.2). Databases were accessed via Oxford University's Bodleian Library subscription. Additional sources were identified through backward citation searching of reference lists from prior systematic reviews (e.g. Zhang, 2024). This approach provided a more thorough overview of the literature and helped reduce the risk of potentially missing important studies not indexed in bibliographic databases (Liberati et al., 2009).

Table 3.2

List of databases

Discipline	Database
Education	ProQuest Social Science Premium Collection (including ERIC); British Education Index
Linguistics	Linguistics Collection (including LLBA)
Psychology	PsycINFO
Multidisciplinary	Web of Science; SCOPUS
Grey literature	ProQuest Dissertations & Theses Global

3.3 Search strategy

When searching for studies, Petticrew and Roberts (2006) highlight the importance of balancing sensitivity (maximising the number of relevant studies found) and specificity (minimising the retrieval of irrelevant ones). To support this process, two librarians from Oxford University's Department of Education were consulted to develop the initial search strategy and Boolean strings for the databases mentioned above (Table 3.2). Piloting the search strategies in Web of Science found that many results focused on Chinese learners of English as a FL. To address this issue and refine the results, broad terms such as 'foreign language', 'Chinese language' and 'Chinese' were removed from the search strings. Terms related to participants' age, including 'primary', 'secondary' and 'tertiary', were initially tested but also later excluded, as the study focused on all CFL learners enrolled in formal institutions. Consequently, two fields of search terms were created to represent the concepts of CFL and Chinese teaching/learning. Terms within each field were connected using the 'OR' function and each field was combined using 'AND'.

The complete Boolean string comprised: ab("Chinese as a foreign language" OR "Mandarin as a foreign language" OR "Chinese character*" OR "Chinese orthograph*" OR "Chinese writing system" OR "Chinese learning") AND ab("Chinese reading" OR "Chinese writing" OR "radical knowledge" OR "Chinese literacy" OR "character recognition" OR "character recall" OR "character acquisition" OR "character knowledge" OR "character retention" OR "character improvement" OR reading development OR "character recall"). Minor wording adjustments were made for certain databases to tailor the search strategy appropriately (see Appendix B for full search strategies for each database consulted).

3.4 Selection process

3.4.1 Initial screening

A second reviewer studying a master's degree in the field of applied linguistics was recruited and briefed on the study's aim/criteria. This person independently screened a random 10% sample of the 1357 abstracts, with their decisions concealed from the main reviewer until both had completed the process. The extent of agreement was calculated using Cohen's Kappa. A Kappa value between 0.81-1.00 (McHugh, 2012) and an inter-rater agreement of 90% or above are generally considered high in education research (Ramezanzadeh & Woore, 2023). The Kappa value was 0.76, with inter-rater agreement of 97%. Although the Kappa value was slightly below the threshold for strong agreement, only four discrepancies (out of 136 abstracts) were identified between the primary and second reviewer. These conflicts were resolved through discussion, during which both reviewers gained a clearer understanding of the reasons behind their disagreements. This process provided confidence that the main reviewer could continue individually screening the remaining abstracts.

Abstracts lacking sufficient information to be confidently ruled out at this stage, along with those that clearly met the eligibility criteria, were included for full-text screening. Screening ceased for abstracts that breached any of the inclusion criteria, and the reviewer noted the reasoning for this decision using 'Exclude 1-9' labels according to the eligibility criteria.

3.4.2 Full-text screening

Bibliographic data of each study in the initial screening process was used to retrieve its full report through the relevant databases or the Bodleian Library. These reports were assessed based on the review's eligibility criteria. A word document was created to record details regarding their inclusion/exclusion outcome.

3.5 Data collection process

Bibliographic data returned by the searches were uploaded to Rayyan on 25 January 2025, an online management tool that allows multiple authors to compare and evaluate studies based on the eligibility criteria (Ouzzani et al., 2016). Bibliographic information was organised in a word document, with notes detailing study characteristics, reasons for including/excluding studies, and relevant findings following the data extraction form (see Section 3.6).

3.6 Data items

After reviewing each report, data were extracted using an adapted version from Chan et al.'s (2022) scoping review of CFL pedagogy, as well as Ramezanzadeh and Woore's (2023) scoping review of teaching and learning second-language (L2) Arabic. Table 3.3 illustrates the data extraction form adopted in this review (see Appendix C for a completed example).

Table 3.3

Data extraction form

Type of information	Subsections
Administrative	Bibliographic information; publication type; language; database source; funding
Contextual	Research location; research question(s); educational level; type of learning and institution; writing script; participant demographics (e.g. age, gender, L1, CFL proficiency, socioeconomic status)
Intervention	Study design and duration; sample size of groups; method of participant allocation; measurement tools; teaching approach(es); teaching/learning materials
Outcomes	Outcome measures; data analysis; effect size; findings; author's conclusions

3.7 RoB assessment

Appraising the methodological quality of individual studies is crucial for identifying possible biases that may compromise the validity and reliability of results (Boland et al., 2014). Biases can occur both within the review process and in the studies being reviewed. A common source of bias in studies is the overestimation of effect size, which can result in overly positive findings, consequently leading to inaccurate conclusions regarding the potential benefits of an intervention (Petticrew & Roberts, 2006). Biases in the review process can additionally stem from authors' actions, flaws in the study design, or lack of transparent reporting in methodology by the reviewers themselves (Gorard, 2024). Certain designs are also considered methodologically more rigorous for identifying causal relationships, such as randomised controlled trials (RCTs), which are less susceptible to allocation bias, compared to quasi-experimental studies (Petticrew & Roberts, 2006). Authors of systematic reviews must assess the overall RoB in each study to ensure that evidence is interpreted accurately.

This review followed Slavin's (1986) best-evidence synthesis approach, which focuses on including a broad range of studies while considering their methodological quality when evaluating the strength and reliability of the assembled evidence. To assess the quality of individual studies, Gorard's (2014) sieve was applied, a popular framework that rates the validity of quantitative outcomes in education intervention research on a scale from zero (indicating a study with high RoB) to four stars (representing low RoB) (see Appendix D for a copy of the sieve and Appendix E for a completed example). The sieve contains six categories: design, scale, dropout, outcomes, fidelity, and validity, with the overall star rating (or evaluation) corresponding to the lowest classification received in any of the categories (Gorard, 2014). If the rating for a particular category cannot be determined from the reports (e.g. because of unclear reporting), that category will automatically be assigned 0*. As there were no guidelines for rating cutoffs in Scale and Dropout, the author consulted their supervisor to develop criteria to ensure consistent scoring:

Scale:

- Trivial: $n \leq 5$
- Very small: $5 < n \leq 15$
- Small: $15 < n \leq 20$
- Medium: $20 < n \leq 50$
- Large: $n \geq 50$

Dropout:

- Minimal: 0-10% dropout (over 90% completion)
- Some: 11-20% dropout (between 80-90% completion)
- Moderate: 21-30% dropout (between 70-80% completion)
- High: Over 30% dropout (less than 70% completion)

Given the subjective nature of quality assessment, the second reviewer involved in the selection process also conducted a blind review of four randomly chosen studies (references were entered into a web-based generator for random selection). There was full agreement on overall quality ratings, though less consistency was found when scoring individual criteria on outcomes, fidelity, and validity. In such cases, both reviewers discussed the discrepancies and successfully reached a consensus. The primary reviewer then proceeded with independently evaluating the remaining full-texts.

3.8 Synthesis methods

A meta-analysis was deemed inappropriate due to the heterogeneity of interventions, comparison groups, and outcome measures across studies. Instead, a narrative synthesis approach was used to summarise the current state of knowledge related to the research questions (Popay et al., 2006). The findings were organised using Petticrew and Robert's (2006) narrative synthesis framework, which consisted of grouping studies into logical categories, examining results within each category, and synthesising insights across all included studies, which involved evaluating the confidence in the accumulated evidence based on RoB assessments.

Chapter 4: Results

This section begins by presenting the findings of the search and selection process, including a flow diagram to illustrate how the final number of studies were identified (see Figure 4.1). It then summarises the general characteristics of the included studies (see Table 4.1), followed by RoB assessments for each experiment. The chapter concludes by assessing the overall strength of the evidence. Study IDs (e.g. S1) are used in Table 4.1 and from Section 4.2.3 onwards to specifically refer to individual studies.

4.1 Study selection

A total of 2875 reports were identified in the initial search, with 1355 abstracts remaining after removing duplicates. From this, 1243 reports failed to meet the eligibility criteria due to wrong study design (n=485), wrong population (n=313), wrong outcomes (n=83), unrelated research (n=306), wrong target language (n=38), and not primary research (n=18), resulting in 112 reports for full-text screening. However, of these, 14 reports from Chinese universities were inaccessible through the Bodleian Library, as these institutions do not participate in Oxford's inter-library loan scheme. Attempts to obtain them by contacting the universities and the principal authors of each study received no response. Eighty reports were further excluded due to mismatches in population (n=20), interventions (n=31), study design (n=5), and outcomes (n=24). In one instance, Google Translate was also used to assess the eligibility of a Spanish-language report. A proficient Spanish speaker was consulted to verify the translation's accuracy, and it was ultimately excluded due to wrong study design. In total, 30 reports met the eligibility criteria and were successfully retrieved for review. The search and selection process is summarised in the PRISMA flow diagram (see Figure 4.1). References of the included studies can be found in Appendix F.

Fig 4.1. *PRISMA flow diagram (adapted from Page et al., 2021)*

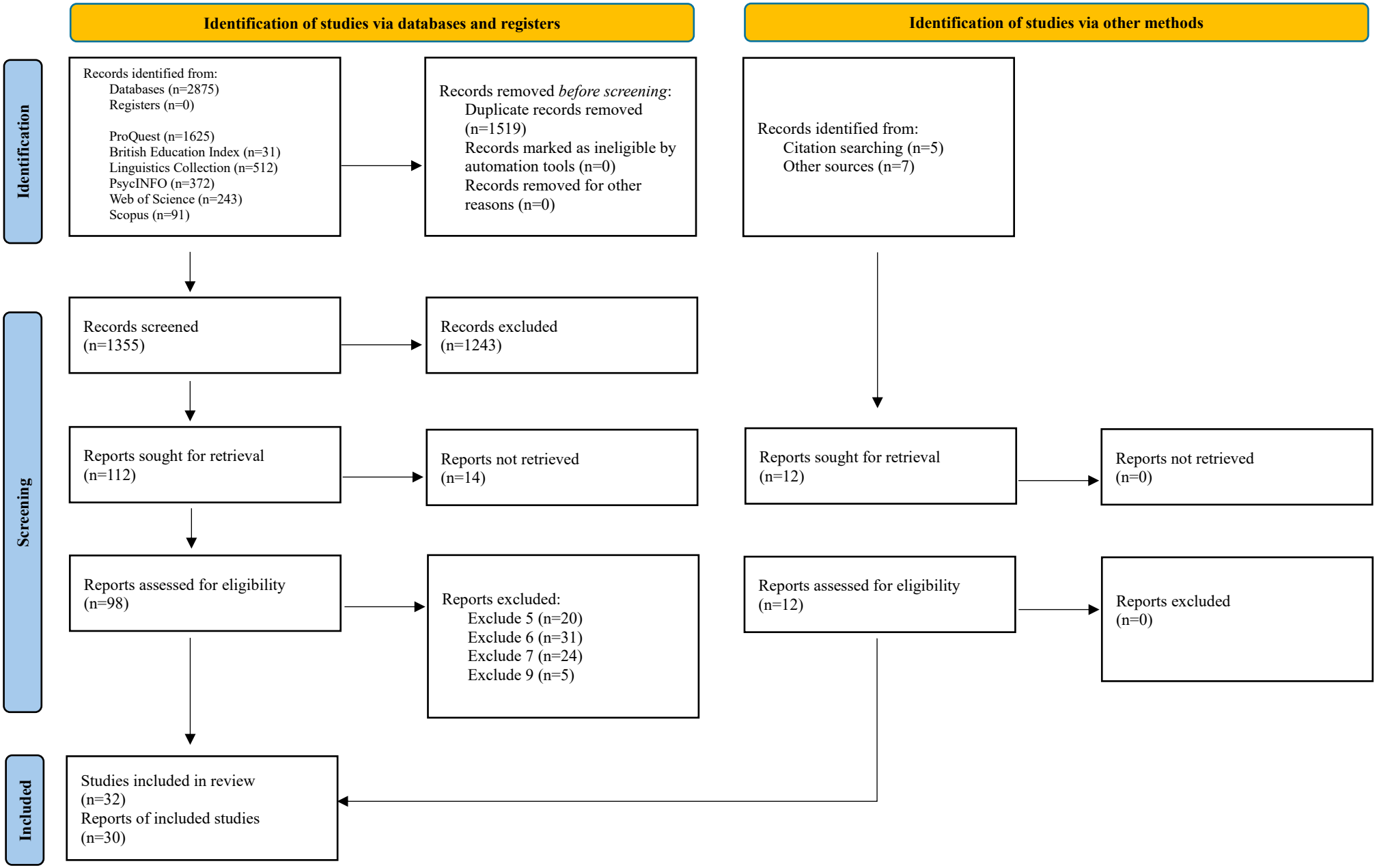


Table 4.1

Study characteristics

Study ID	Location	Education level	Design	Treatment	Comparator (or control group)	Writing script	Sample size	Length	Assessment focus	Results
1. Chang et al. (2014)	USA	University	Quasi-experimental crossover (pre-post- and delayed post-test)	Chunking	(a) Pencil-and-paper handwriting (b) Passive reading (c) Stroke-order writing	Simplified	41	4 days	Character-recall and retention	Favoured treatment group
2. Chen (2020)	Thailand	Secondary	Single-group (pre-post-test)	Interactive learning platform	None	Simplified	12	4 weeks	Reading	Favoured treatment group
3. Chiu (2024)	USA	Secondary	Quasi-experimental crossover (pre-post-test)	Typing with pinyin input	Pencil-and-paper handwriting	Simplified	23	5 weeks	Character knowledge, vocabulary	Favoured treatment group
4. Chung (2007)	Australia	Secondary	RCT crossover (pre-post- and delayed post-test)	Study 1: Flashcards (various ways of presenting character information)	Study 2: Flashcards (with and without colour-coding)	Simplified	32 in Study 1, 32 in Study 2	2-3 lessons per week (until pupils learned all characters)	Meaning-recall and retention	Favoured flashcards in both studies
5. Harvey & Brooks (2022)	USA	Primary	Quasi-experimental (pre-post-test)	Typing in chat-groups	Pencil-and-paper handwriting	Simplified	56	8 weeks	Reading	No significant differences between groups
6. He & Huang (2014)	Australia	University	RCT (pre-post-test)	(a) Digital materials with English meaning and animated etymological information (b) Paper materials with English meaning and etymological information via images	Paper materials with English meaning and character	Simplified	21	1x 30-min lesson	Character-recognition	Favoured digital materials group

7. Ke & Dubravac (2021)	USA	University	Quasi-experimental (mid-post- and delayed post-test)	Early exposure to characters	Delayed exposure to characters	Simplified	2	4 hours per week over 14 weeks	Character recognition, retention	Favoured treatment group
8. Knell & West (2017)	USA	Secondary	RCT (post-test)	Early exposure to characters	Delayed exposure to characters	Traditional	48	1 academic year	Character-recognition, reading	Favoured treatment group on one measure
9. Li (2004)	USA	University	Quasi-experimental (pre-post-test)	Grouping characters by shared radicals	Rote memorisation	Simplified	50	8 weeks	Character-recognition, reading	Favoured treatment group
10. Li & Tong (2020)	USA	Primary	RCT (pre-post- and delayed post-test)	Verbal-coding* with etymological information	Verbal-coding* with pictures	Traditional	100	19 weeks	Character-recognition, retention	Favoured treatment group
11. Osborne (2016)	Ireland	Secondary	Quasi-experimental (mid-post-test)	(a) Rote learning (b) Delayed character instruction (c) Colour-coding characters	Unity curriculum (focusing on listening, reading, writing, speaking)	Not specified	98	1 academic year	Character-recognition and recall, reading, pinyin spelling	Effects differed across measures
12. Osborne (2018)	Ireland	Secondary	RCT (pre-post-test)	(a) Rote learning (b) Delayed character instruction (c) Colour-coding characters	Unity curriculum (focusing on listening, reading, writing, speaking)	Not specified	85	16 weeks	Character-recognition and recall, pinyin spelling	Effects differed across measures
13. Osborne et al. (2020)	Ireland	Secondary	Quasi-experimental (mid-post-test)	(a) Rote learning (b) Delayed character instruction (c) Colour-coding characters	Unity curriculum (focusing on listening, reading, writing, speaking)	Simplified	98	8 weeks	Character-recognition and recall, reading, pinyin spelling	Effects differed across measures
14. Osborne et al. (2022)	Ireland	Secondary	Quasi-experimental (mid-post-test)	(a) Rote learning (b) Delayed character instruction (c) Colour-coding characters	Unity curriculum (focusing on listening, reading, writing, speaking)	Simplified	80	28 weeks	Character-recognition and recall, reading, pinyin spelling	Effects differed across measures

15. Packard (1990)	USA	University	Quasi-experimental (post-test)	Early exposure to characters	Delayed exposure to characters	Traditional	23	26 weeks	Character production	Favoured 'delayed' group
16. Poole & Sung (2015)	USA	University	Quasi-experimental (post-test)	(a) Pinyin-only (b) Passive reading	Pencil-and-paper handwriting	Not specified	9	4x 30-min lessons	Character-recognition, speaking	Effects differed across measures
17. Poole et al. (2022)	USA	Primary	Single-group (pre-post-test)	Digital game	None	Simplified	32	4x 25-min lessons over 4 weeks	Character recognition, reading	Favoured treatment group
18. Ren (2004)	Australia	Secondary	Cluster RCT (post-test)	OVAL-writing**	PowerPoint	Not specified	19	10 weeks	Character retention	Favoured treatment group
19. Shen (2010)	USA	University	Single-group (post- and delayed post-test)	Study 1: Verbal-coding with images Study 2: Verbal-coding only	None	Simplified	40 in Study 1, 45 in Study 2	12 lessons	Concrete and abstract words	Effects differed across measures
20. Taft & Chung (1999)	Australia	University	Quasi-experimental (post- and delayed post-test)	Radicals introduced at various times: (a) Before the first lesson (b) Early in the first lesson (c) After the third lesson	No radicals were introduced	Simplified	40	3x 15-min lessons over one week	Character-recall, and meaning-pairings	Favoured 'radicals early' group
21. Tsai (2014)	USA	University	Quasi-experimental crossover (post-test)	Character information with audio: (a) Digital writing without feedback (b) Typing input only (c) Digital writing with feedback	Pencil-and-paper handwriting	Simplified	63	10 lessons over 16 weeks	Character-recognition and production	Favoured control group
22. Wang (2005)	USA	University	Cluster RCT (post- and delayed post-test)	(a) Text plus animation aids (b) Text with audio (c) Text plus animation and audio	Printed text	Traditional	72	1x 50-min lesson	Character-recognition, recall, and retention	Favoured text plus animation and audio group

23. Wang (2024)	USA	University	Cluster RCT (pre-post-test)	Digital game	PowerPoint	Simplified	49	4x 20-min lessons over 4 weeks	Radical knowledge	Favoured treatment group
24. Wu (2012)	USA	University	RCT (pre-post-test)	Text plus audio	Printed text	Simplified	65	4 weeks	Character-recognition and recall, reading	Favoured treatment group
25. Xu & Jen (2005)	USA	University	Cluster RCT (mid-post-test)	Typing with pinyin input	Pencil-and-paper handwriting	Not specified	138	3 academic years	Character-recognition	Favoured treatment group
26. Xu & Ke (2020)	USA	University	RCT (pre-post- and delayed post-test)	Imitate character strokes through body movement	Viewing stroke-order sequences of characters onscreen	Simplified	53	1x 40-min lesson	Character-recognition, recall, and retention	Favoured treatment group
27. Xu & Padilla (2013)	USA	Secondary	Matched-pairs RCT (pre-post- and delayed post-test)	Meaningful interpretation and chunking	Rote repetition and stroke-order writing	Simplified	108	4 days	Character-recognition, recall, and retention	Favoured treatment group
28. Xu et al. (2013)	USA	University	RCT crossover (pre-post- and delayed post-test)	(a) Digital writing (b) Animated demonstrations of stroke-order writing	Passive reading (online)	Traditional	36	6x 20-40-min lessons	Character-meaning, recall, and retention	Favoured digital writing group
29. Xu et al. (2014)	USA	University	Quasi-experimental (pre-post- and delayed post-test)	Grouping characters by shared radicals	Learning characters across eight radical groups	Simplified	88	4 lessons	Character and radical knowledge	Favoured treatment group
30. Zhang (2019)	Australia	Primary	Single-group (pre-post-test)	Online comics	None	Simplified	60	10 weeks	Character production	Favoured treatment group

Note. *Verbal encoding involved the instructor teaching characters through verbal definitions and/or explanations.

**OVAL-writing approach: observe, visualise, articulate, listen, write. The instructional sequence involved students observing the character's structure, mentally visualising its form, practising pronunciation, listening to the instructor's pronunciation, and concluding with handwriting practice.

4.2 Study characteristics

Table 4.1 outlines the general characteristics of each included study. The subsections below examine these aspects in greater detail, highlighting trends in publication type, location, educational level, research design, writing script, instructional approaches, participant demographics, and outcome measures.

4.2.1 Publication details

Figure 4.2 displays the distribution of included studies across five-year publication intervals, spanning 1990-2024. The earliest study was published in 1990, with increases in research output from 2020 onwards. Following 1990, there has been a gradual yet infrequent rise in publications, with peaks in 2014 and 2020 (n=4 for each year). Table 4.2 illustrates that 23 out of 30 studies were peer-reviewed journal articles, followed by doctoral (n=4) and master's dissertations (n=2), with only one conference proceeding.

Fig. 4.2

Included studies by publication year

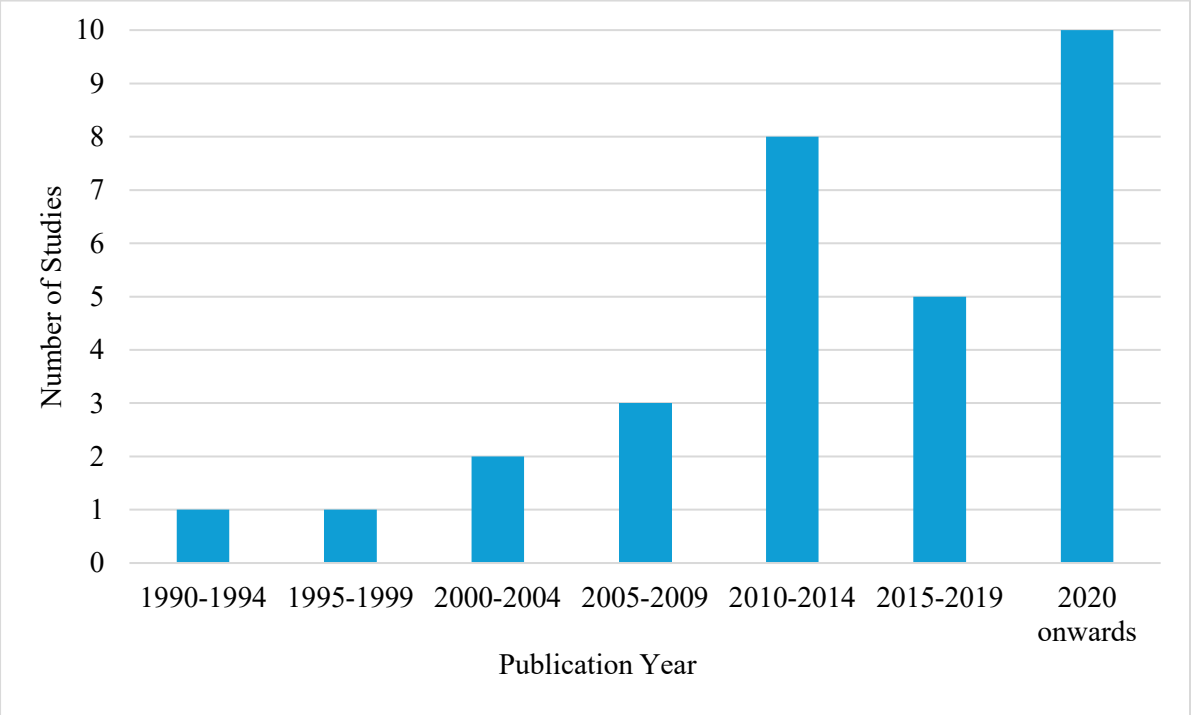


Table 4.2

Included studies by publication type

Publication Type	Quantity
Conference paper	1
Doctoral thesis	4
Journal article	23
Master's dissertation	2

4.2.2 Location and educational level

Most studies included in this review were conducted in the United States (n=20), followed by Australia (n=5), Ireland (n=4), and Thailand (n=1). Sixteen studies recruited university subjects, with 14 of these taking place in American institutions and the remaining two in Australia. Ten studies explored outcomes in secondary education, while four examined orthographic learning in primary schools. This information is depicted in Figures 4.3 and 4.4.

Fig. 4.3. *Educational level*

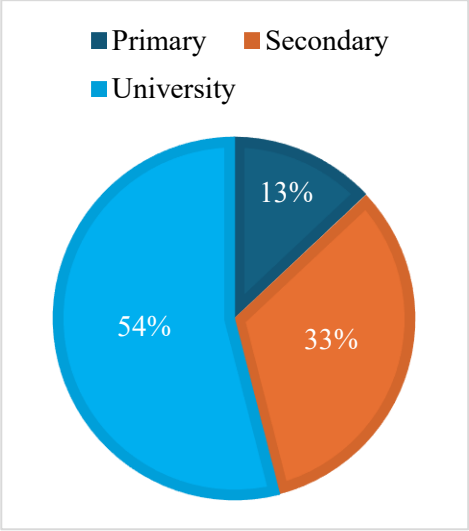
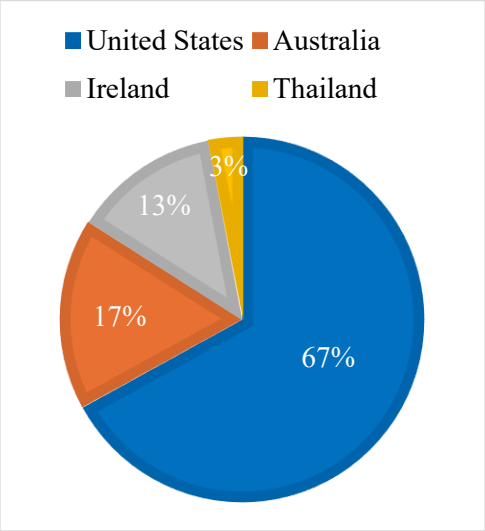


Fig. 4.4. *Study location*



4.2.3 Instructional approaches

A wide range of pedagogic methods were reported across the studies. Many investigated the use of CALL through diverse formats, including digital games (S2, S17, S23), chat-based texting (S5), animated demonstrations of character components with audio (S21, S22, S24, S28), online comics (S30), and kinaesthetic-haptic learning combining body movement with visual prompts (S26). These were typically compared against conventional techniques, such as handwriting, pencil-and-paper instruction, rote memorisation, and repetition. However, the

dominant trend across studies was to explore different methods of presenting character components – for example, teaching new characters using etymological information (S6, S10), grouping shared radicals to form compound characters (S1, S9, S20, S27, S29), and colour-coding tone marks, characters, or pinyin (S4, S11, S12, S13, S14). Other studies examined the optimal timing of orthographic learning, comparing delayed instruction – where characters were taught after an initial period focused solely on pinyin – with early instruction, where characters were introduced from the beginning of the intervention (S7, S8, S15). Table 4.3 displays the breakdown of pedagogic methods.

4.2.4 Study design and writing script

Research designs were evenly distributed between RCTs (n=13) and quasi-experimental studies (n=13), the latter of which did not involve randomly assigning subjects to conditions. Among the RCTs, some employed cluster randomisation (S18, S22, S23, S25), meaning that intact classes were randomly allocated to treatment or control groups. Two studies used RCT crossover design (S4, S28), where participants experienced two or more interventions in a randomised sequence, with different treatments delivered at different time-points. One study (S27) incorporated pair-matching – that is, matching subjects with similar scores on pre-test results and randomly assigning one from each pair to the treatment or control condition. In the remaining RCTs (S6, S8, S10, S12, S24, S26), subjects were individually randomised into groups. These studies followed a parallel design, such that one group received only one version of the study's intervention. Four studies (S2, S17, S19, S30) adopted a single-group design, investigating the intervention's effects using one group of participants (i.e. no comparator).

Although most studies assessed intervention effects by comparing pre-post- and/or delayed post-test outcomes across groups, 12 out of 30 studies only implemented post-test measures. Specifically, six studies evaluated outcomes mid-way through the intervention (e.g. after five weeks of instruction) and upon its completion, while another six studies relied solely on post- and/or delayed post-test data. Twenty studies also explored the teaching and learning of simplified characters, while only four focused on traditional characters. Six studies did not clarify which writing script was used. Such reports lacked information for the reviewer to make an informed judgement (i.e. no appendix or examples within the study were provided to determine which script was taught). Table 4.4 outlines the research designs.

Table 4.3

Included studies by instructional approach

Instructional category	Specific method	Quantity	Study ID*
CALL	Digital games	2	S17, S23
	Chat-based texting	1	S5
	Animated orthographic input with audio	4	S6, S21, S22, S28
	Online comics	1	S30
	Kinaesthetic-haptic learning with images	1	S26
	Typing characters with pinyin input	3	S3, S21, S25
	Animated demonstrations of stroke-order	1	S28
	Interactive learning platform	1	S2
	Printed text plus audio	1	S24
	Digital writing	1	S28
Conventional techniques	Pencil-and-paper instruction (including handwriting, stroke-order writing)	9	S1, S3, S5, S6, S16, S21, S25, S27, S28
	Rote memorisation/repetition	7	S9, S11, S12, S13, S14, S17, S27
	PowerPoint	2	S18, S23
	Passive reading	5	S1, S16, S22, S24, S28
	Pinyin only	1	S16
	Verbal definitions of characters	1	S19
Character components	Grouping the same radicals to form compound characters	3	S9, S20
	Colour-coding characters and/or tone marks	5	S4, S11, S12, S13, S14
	Visual chunking	3	S1, S27, S29
	Flashcards (different sequences of information presented e.g. pinyin-character-English meaning vs. character-pinyin-English meaning)	1	S4
	Etymological explanations of radicals	2	S6, S10
	Verbal definitions of characters with pictures	1	S19
	Learning characters across eight different radical groups	1	S29
Timing of instruction	Early vs. delayed character introduction	3	S7, S8, S15
Holistic approach	Unity curriculum (balancing speaking, reading, writing, and listening)	4	S11, S12, S13, S14
	OVAL-writing	1	S18

Note. *Study IDs are listed more than once in cases where authors evaluated multiple instructional approaches within a single study.

Table 4.4

Summary of research designs

Research Design	Assessment Timing	Quantity	Study ID
Cluster RCT	Pre-post-test	1	S23
	Post-test only	1	S18
	Post- and delayed post-test	1	S22
	Mid-post-test	1	S25
Crossover RCT	Pre-post- and delayed post-test	2	S4, S28
Individual RCT	Pre-post-test	3	S6, S12, S24
	Pre-post- and delayed post-test	2	S10, S26
	Mid-post-test	1	S8
Matched-pair RCT	Pre-post- and delayed post-test	1	S27
Quasi-experimental	Pre-post-test	3	S3, S5, S9
	Pre-post- and delayed post-test	2	S1, S29
	Mid-post- and delayed post-test	1	S7
	Mid-post-test only	3	S11, S13, S14
	Post-test only	2	S15, S16
Quasi-experimental crossover	Post-test only	1	S21
	Post- and delayed post-test	1	S20
Single-group	Pre-post-test	3	S2, S17, S30
	Post- and delayed post-test	1	S19

4.2.5 Sample size and duration

The number of participants per study varied widely, ranging from two in S7 where subjects opted in voluntarily, to 138 in S25 which gathered data over a three-year period. As shown in Table 4.5, 27 out of 30 studies recruited fewer than 100 participants in total, potentially limiting statistical precision. All interventions took place in group settings, whether as single-, treatment or control groups, except in S4, S7, and S26, where participants received individual instruction from the principal author of each study. The smallest group sample size appeared in S16, with only three subjects from intact classes assigned to one condition. In contrast, the largest group size was observed in S27, with 53 students in the experimental group and 55 in the control group. Although S25 reported a total of 138 subjects – the largest sample among the included studies – the methodology lacks details regarding how many participants were distributed across conditions over the three-year data collection period.

The duration of interventions fluctuated greatly, from a single 30-minute lesson (S6) to treatments spanning 1-3 academic years (S8, S11, S14, S15, S25) (see Table 4.6). Further, the intensity of interventions differed substantially across the more longitudinal studies: 80-minute sessions over 22.5 weeks (S8), two one-hour classes vs. eight hours per week for one

year (S11, S15), and 56 hours of teaching exposure over 26 weeks (S14). While the total hours of instruction in S25 were not specified, each intervention lasted one academic year (two semesters), with data collected from three consecutive cohorts over three years (six semesters altogether). Moreover, all groups within a given study did not receive the same amount of treatment exposure (e.g. the 'early' group in S8 received 30 hours of orthographic instruction beginning in Week 2, while the 'delayed' group received 18-20 hours where character teaching started three months later). Such discrepancies may compromise the interpretation of results.

Table 4.5

Sample size

Total Number of Participants	Quantity	Study ID
0-20	3	S2, S7, S16, S18
21-40	7	S3, S4, S6, S15, S17, S19*, S20, S28
41-60	8	S1, S8, S9, S19*, S23, S26, S30
61-80	5	S14, S21, S22, S24
81-100	5	S10, S11, S12, S13, S29
101-120	1	S27
121-140	1	S25

Note. *40 in Experiment 1 of S19 and 45 in Experiment 2.

Table 4.6

Study duration

Duration	Quantity	Study ID
Less than 1 week	5	S1, S6, S22, S26, S27
1-4 weeks	5	S2, S17, S20, S23, S24
5-8 weeks	4	S3, S5, S9, S13
9-12 weeks	2	S18, S30
13-16 weeks	3	S7, S12, S21
17-20 weeks	1	S10
20+ weeks	5	S8, S11, S14, S15, S25
Unclear*	5	S4, S16, S19, S28, S29

Note. *Reports lacked sufficient information regarding study duration (e.g. while S19 stated that students completed 12 intervention lessons, the author does not specify the length of each lesson or how often the lessons took place).

4.2.6 Participant demographics

4.2.6.1 CFL proficiency

Descriptions of CFL proficiency were often vague, using terms like ‘beginner’, ‘novice’ or ‘non-fluent’. In most cases, proficiency was determined by researchers, language teachers, or through students’ self-assessments based on the total number of characters they could recognise prior to the intervention. In eight university-based studies, pupils were described as enrolled in first-year beginner classes (S1, S9, S19, S23, S24, S25, S28, S29). However, the criteria for classifying learners as ‘beginners’ also varied significantly. In S1 and S29, for example, beginners had already learned around 180 characters and were familiar with pinyin and stroke-order writing. Conversely, S24 categorised pupils as beginners despite already acquiring approximately 400 characters. S8 also measured proficiency based on the total hours of instructional exposure, with students considered beginners after completing roughly 150 hours of Chinese classes. Finally, some studies reported no prior exposure to Mandarin (S7, S8, S11, S12, S13, S16).

While the Hanyu Shuiping Kaoshi (HSK) exists as a standardised proficiency framework for CFL learners, it is not commonly used across international research contexts. Among the studies reviewed, only S14 explicitly referenced the HSK, classifying participants at Level 1 (i.e. complete beginners). In three cases (S5, S17, S30), subjects were enrolled in bilingual immersion programs (English-Chinese) in primary-school settings – that is, students in S5 had completed at least three years of immersion education and had five years of formal CFL instruction; S30 recruited subjects from a Content and Language Integrated Learning (CLIL) context who had been studying Chinese since kindergarten (with instruction divided evenly between English and Chinese; CLIL is an educational approach to bilingual learning where subject matter is taught through the L2/FL with the explicit intention of developing both); and, S17 did not specify how long students had been studying in bilingual programs. S29 compared beginner and intermediate learners, with the latter group reportedly having acquired around 530 characters compared to 180 among the beginners. These findings highlight a general lack of consistency in the classification and reporting of CFL learner proficiency across studies.

4.2.6.2 Gender and L1 background

Twelve reports did not state participants' gender, while three studies referred to them as mixed-gender without specifying female-to-male ratio. S4 and S18 only recruited males, both conducted in secondary contexts where pupils were enrolled in single-sex private schools. Only S3 acknowledged gender beyond the male-female binary by noting the participation of a student who identified as non-binary – this was the sole instance of gender diversity being explicitly mentioned. Among the remaining studies, some reported higher proportions of female subjects (S1, S10, S19, S21, S22, S26), others indicated greater male participants (S3, S9, S24, S27), and two reported relatively balanced gender distribution (S6, S7).

Eight studies did not provide specific information regarding participants' L1 background (S8, S9, S10, S15, S17, S21, S25, S26). However, it should be highlighted that almost all studies were conducted in geographic regions where English is the majority language (i.e. United States, Australia, Ireland). S2 identified Thai as students' L1, whereas S6 described students as having non-logographic L1 backgrounds. S23 also reported two participants with L1 Vietnamese. The remaining nineteen cases stated English as learners' L1.

4.2.7 General outcomes

All included studies implemented quantitative measures of character acquisition and/or reading performance, as outlined by the eligibility criteria. Twenty-eight reports concentrated on orthographic knowledge, which encompassed a range of subskills including character or radical recognition, recall, retention, and meaning recall. Two studies focused on reading only. Assessments involved a range of receptive and productive tasks, such as re-ordering Chinese sentences within short paragraphs/conversations, matching characters/words to images, and translating Chinese to English and vice versa. These diverse outcome measures reflect the field's emphasis on early character-level processing as a foundation for developing broader vocabulary knowledge and CFL proficiency.

4.3 RoB

Table 4.7 outlines the quality ratings assigned to each study using Gorard's (2014) sieve. Ratings were categorised between 0-4* (studies rated 0-1* are classified as weak with high RoB, 2* as moderate, and 3-4* as strong with low RoB). While the final corpus comprises 30 reports, S4 and S19 conducted two distinct experiments, which were treated as separate units of analysis, bringing the total number of analysed experiments to 32.

Table 4.7

RoB assessment

Study	Design	Scale	Dropout	Outcomes	Fidelity	Validity	Rating
Chang et al. (2014)	2	2	0	2	3	4	0*
Chen (2020)	0	1	4	4	3	3	0*
Chiu (2024)	2	1	4	4	4	3	1*
Chung (2007) (Experiment 1)	4	3	0	1	4	3	0*
Chung (2007) (Experiment 2)	4	3	0	1	4	3	0*
Harvey & Brooks (2022)	2	3	3	4	3	3	2*
He & Huang (2014)	4	1	4	2	4	3	1*
Ke & Dubravac (2021)	2	0	4	3	4	1	0*
Knell & West (2017)	4	3	3	2	3	2	2*
Li (2004)	2	3	4	1	3	3	1*
Li & Tong (2020)	4	4	0	2	4	3	0*
Osborne (2016)	2	3	0	1	4	2	0*
Osborne (2018)	4	3	0	1	4	3	0*
Osborne et al. (2020)	2	3	0	1	4	2	0*

Osborne et al. (2022)	2	3	4	1	4	3	1*
Packard (1990)	2	1	4	1	3	2	1*
Poole & Sung (2015)	2	0	0	1	4	2	0*
Poole et al. (2022)	0	2	2	2	3	3	0*
Ren (2004)	4	0	2	1	2	2	0*
Shen (2010) (Experiment 1)	0	3	3	1	4	2	0*
Shen (2010) (Experiment 2)	0	3	3	1	4	2	0*
Taft & Chung (1999)	2	1	0	2	3	2	0*
Tsai (2014)	1	4	3	1	3	4	1*
Wang (2005)	4	0	4	3	4	2	0*
Wang (2024)	4	0	4	2	3	2	0*
Wu (2012)	4	3	3	3	4	3	3*
Xu & Jen (2005)	1	0	0	1	3	2	0*
Xu & Ke (2020)	4	4	2	2	4	3	2*
Xu & Padilla (2013)	4	4	3	4	4	3	3*
Xu et al. (2013)	3	3	4	2	4	4	2*
Xu et al. (2014)	3	3	4	3	3	3	3*
Zhang (2019)	0	4	4	2	2	3	0*

4.3.1 Scale/selection bias

Six studies were rated as having high risk of selection bias (0-1*), primarily due to limited sample size: only two participants volunteered in S7, where students were allowed to choose their preferred teaching approach (early or delayed character instruction); S16 recruited only nine subjects, with three students assigned to one experimental condition each; and, as the remaining studies (S18, S22, S23, S25) adopted cluster randomisation with only one class assigned per condition, the number of independent comparison units resulted in one case per group. Seventeen studies demonstrated low risk (3-4*), whereas seven exhibited moderate risk linked to low completion rates (70-80%; 2*). It is noteworthy that 22 out of 30 reports recruited students who had chosen to enrol in Chinese classes at their respective institutions, indicating some degree of self-selection bias, as pupils were likely motivated to learn.

4.3.2 Study design

Random allocation of participants to treatment or control conditions is widely considered the most robust method for establishing baseline equivalence prior to the intervention. On this basis, 12 RCTs in the final corpus received 4* for strong design. However, none of the studies described the randomisation process, whether authors used statistical software programs, a web-generated random sequence, or manual allocation methods, limiting transparency and replicability. Eighteen of the remaining studies were quasi-experimental or matched based on previous scores, with four of the 18 adopting single-group designs. These studies were rated between 0-3* with single-group interventions receiving 0* due to lack of comparator.

4.3.3 Outcomes

Fifteen studies offered little to no details regarding how outcomes were internally validated or how assessment measures were reliable instruments, receiving 0-1* as tests were unclear and authors did not report internal validity/inter-rater reliability. Despite four studies using standardised measures (rated 4*), the respective authors did not explain or offer details about the marking criteria or whether raters were blind to participant allocation. Ten studies obtained a 2* rating, as authors often combined validated assessment tools in some instances with self-designed rubrics in others.

4.3.4 Dropout

Twenty-two studies (from 20 reports) provided complete data, clearly documenting attrition and reasons for participant withdrawal. Three studies reported moderate attrition rates (21-30%), attributed to incomplete data (S19), unexpected student absences (S28), and the subsequent exclusion of participants with heritage language backgrounds whose inclusion could have compromised results (S18). The remaining 10 studies did not report on attrition.

4.3.5 Validity and fidelity

While it is acknowledged that controlling for all confounding variables is difficult without randomisation, the reviewer specifically focused on controlling known differences between research groups in these categories of Gorard's (2014) sieve. It was expected that authors would report basic demographic information, including age, gender, socioeconomic status, and L1 background. However, such data were regularly omitted, which would have resulted in many studies receiving 0-1*, irrespective of whether authors had controlled for other variables like CFL proficiency. Therefore, studies in this category were rated 1-2* when pre-tests were not administered or when participants did not follow the intervention properly. Fidelity refers to how clearly and consistently an intervention was delivered: 18 studies obtained 4* with no indication of instructor differences; 12 experiments earned 3* due to variations in teacher; and two achieved 2* because the intervention was not adequately described to enable replication.

4.3.6 Cumulative confidence across studies

Altogether, 19 studies received 0*, five were rated 1*, four attained 2*, and three achieved 3*. No study obtained the maximum score of 4*. Table 4.8 illustrates that studies with high RoB comprised two thirds of included papers. Problems arose primarily in using appropriate measurements, such as standardised/validated instruments, accounting for confounders in the design and data analysis, and limited sample sizes when cluster RCT designs were employed. Figure 4.5 shows that all studies reported positive gains except for one, which found neutral results on reading comprehension. A critical factor in interpreting the effectiveness of the intervention is the nature of the comparator used in these studies: 17 compared the treatment approach to traditional methods, such as handwriting, repetition, rote learning, passive reading, and presenting information through PowerPoint slides. Four studies lacked baseline comparisons: three focused on CALL, while one conducted two single-group experiments to compare verbal definitions of characters with and without pictorial support. The remaining

nine studies compared different instructional approaches, potentially providing greater insights than comparisons to either no instruction or more conventional techniques (e.g. early vs. delayed character exposure, colour-coding vs. delayed instruction vs. communicative language teaching, different timelines for introducing radical information, imitating character strokes through body movements vs. viewing stroke-order animations online, and grouping characters by shared radicals vs. learning across different radical groups). Despite positive results, the cumulative confidence across studies is limited, largely due to the paucity of robust comparative designs and methodological weaknesses (explored further in Chapter 5).

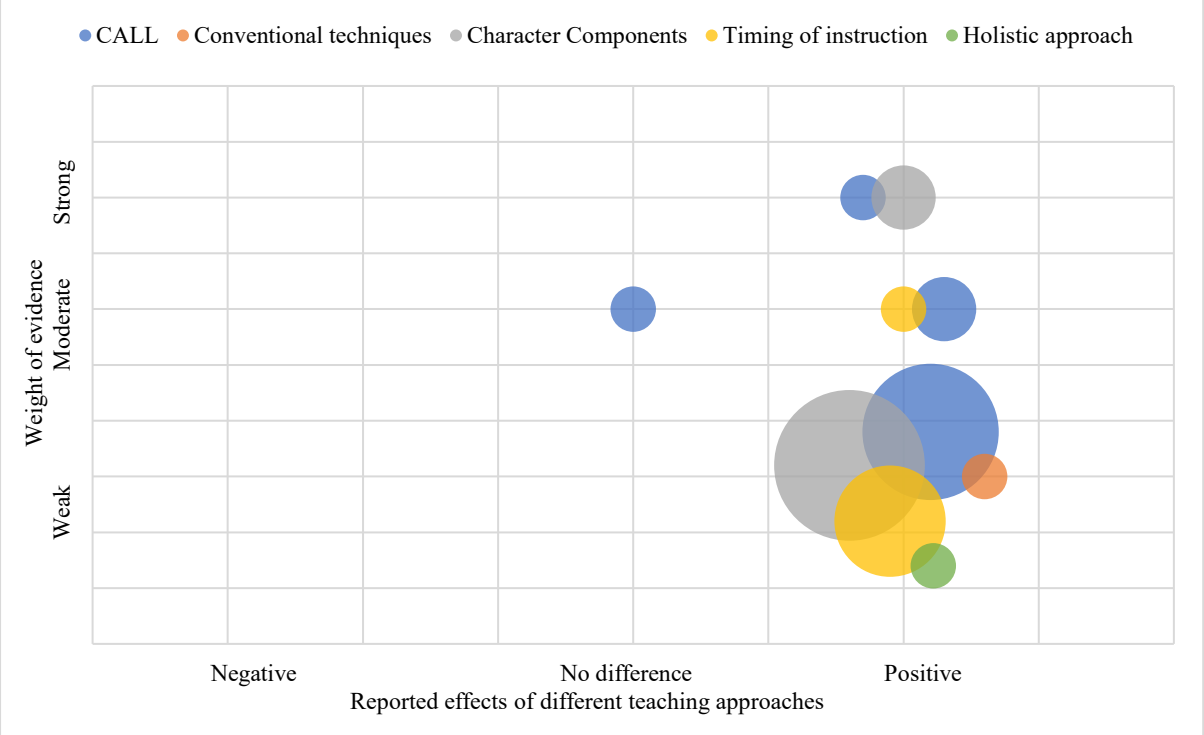
Table 4.8

Strength of evidence and distribution of effects

Strength of evidence	Positive	Unclear/Mixed	Neutral
4*	–	–	–
3*	S24, S27, S29	–	–
2*	S8, S26, S28	–	S5
1*	S3, S6, S9, S14, S15	–	–
0*	S1, S2, S4, S7, S10, S11, S12, S13, S16, S17, S18, S19, S20, S21, S22, S23, S25, S30	–	–

Fig. 4.5

Reported effects of different teaching approaches



Chapter 5: Synthesis

While interest in CFL has grown in certain geographic regions, and various instructional strategies have been investigated in both peer- and non-peer-reviewed literature, there is minimal reliable evidence promoting the efficacy of any one approach in enhancing linguistic outcomes. Previous reviews (e.g. Li, 2020; Zhang, 2024) feature incomplete reporting, focus narrowly on L1 English CFL learners, and lack RoB assessments, but still conclude that all strategies examined are effective. This review reveals that the methodological quality of the included studies was weak, restricting the ability to provide strong pedagogical suggestions. Nonetheless, these findings offer insight into how CFL teaching has been examined in formal settings and highlights the need for future explorations to employ more rigorous study designs. The following sections are structured around the review's research questions.

5.1 What pedagogical approaches to teaching Chinese orthography have been evaluated for effectiveness among school- and university-aged CFL learners?

The review identified five main pedagogical approaches: CALL; conventional techniques; character component instruction; timing of orthographic exposure; and holistic methods. CALL studies varied in interactivity and learner engagement, with more passive approaches using animated stroke-order demonstrations accompanied by audio to enhance form-meaning connections. Conversely, interactive CALL included formats that prompted text production (e.g. online comic creation, digital games), possibly promoting deeper cognitive processing. Several studies also explored different ways of presenting single characters/words. Popular methods involved grouping characters by shared radicals, offering etymological orthographic input, introducing flashcards, and colour-coding pinyin/characters. Other studies examined the optimal timing for teaching orthography, comparing immediate exposure with a delayed approach where students exclusively learned pinyin before transitioning to characters. Conventional (or traditional) strategies were largely rooted in L1 Chinese teaching, emphasising repetition, handwriting, and rote learning. These were often compared against other approaches like CALL. Meanwhile, holistic methods balancing speaking, listening, reading, and writing were used as comparators against explicit orthographic instruction, focusing primarily on communicative language teaching. One study investigated the effects of an OVAL-writing technique, which combined visualisation, oral articulation, handwriting, and auditory input from the teacher. As this review aims to present practical strategies for CFL educators, summaries of instructional methods are provided below. Participant and design details were outlined in Chapter 4.

5.2 What are the effects of these methods on CFL learners' character acquisition and reading performance?

5.2.1 Timing of orthographic exposure

Three studies compared two methods to this approach: an 'early' group learned characters from the beginning; and a 'delayed' group initially concentrated on pinyin (e.g. for four weeks) with characters introduced later in the intervention (e.g. from Week 5). Knell and West (2017) observed that immediate exposure facilitated better reading comprehension, though no significant between-group differences emerged in character recognition. Ke and Dubravac (2021) identified significant advantages in character recognition for students exposed early rather than later. Conversely, Packard (1990) reported significant benefits in pinyin transcriptions and oral proficiency for delayed instruction only. While Osborne (2016, 2018) and Osborne et al. (2020, 2022) did not directly compare these approaches, they found significant gains in pinyin knowledge with delayed exposure. This was evaluated against rote memorisation, colour-coding, and a holistic approach integrating listening, speaking, reading, and writing (see Section 5.2.3.2). Collectively, these studies indicate mixed results regarding the optimal timing for orthographic learning.

However, a notable limitation in these studies is the unequal allocation of teaching time across groups. Ke and Dubravac (2021) conducted a 14-week intervention, with the early group starting in Week 1 and the delayed condition in Week 5. Packard (1990) compensated the delayed group with three additional one-hour lessons per semester. Osborne's research does not specify when character teaching began post-delay. These inconsistencies affect the validity of results, as differences in outcomes may reflect variations in exposure rather than instructional timing. Study quality does little to moderate these effects: despite Knell and West (2017) obtaining the highest rating (2*) for trustworthiness among these studies, the early group were introduced to characters in Week 2 (around 30 contact hours), while their counterparts began three months later (totalling 18-20 hours). It is also noteworthy that Ke and Dubravac (2021; rated 0*) recruited only two participants, both of whom self-selected into groups and were described as highly motivated learners, possibly reflecting selection bias. The authors do not address confounding variables like motivation or out-of-school contact, which may have influenced outcomes as students completed weekly quizzes, most likely engaging in independent study. There is insufficient evidence to deduce whether early or delayed exposure is more successful for CFL development.

5.2.2 CALL instruction

5.2.2.1 Digital games

Two studies employed this method. Poole et al. (2022) evaluated the effects of an online game, where CFL learners completed linguistic quests across China as adventurers. In this single-group study, participants used in-game glossaries to understand content and complete missions independently. The game was supplemented with workbook exercises, including sentence construction, handwriting practice, and character-picture matching. Findings showed significant gains in both vocabulary and reading skills from pre- to post-test. Completion of workbook exercises was significantly correlated with improvements in reading performance.

Wang (2024) designed a digital escape-room game using Google Forms to teach radical awareness, comparing its effectiveness to traditional orthographic instruction delivered via PowerPoint slides. Participants completed activities involving semantic radical identification and compound character decoding. Game progression depended on correctly identifying radical functions through multiple-choice questions. Compared to their peers, digital game-based instruction achieved significantly greater gains in radical knowledge from pre- to post-test. While both of the preceding studies showed positive findings, Poole et al. (2022) lacked a control group and Wang's (2024) cluster RCT involved only two cases, severely limiting its statistical power. Therefore, meaningful conclusions about the effectiveness of digital games cannot be drawn, as research designs make it unclear whether anything more than exposure and time were responsible for increased orthographic knowledge.

5.2.2.2 Texting, interactive learning platform, online comics

Harvey and Brooks (2022) compared chat-based texting with pencil-and-paper handwriting to evaluate its impact on reading and character production. The texting condition engaged in online discussions on everyday topics using prompts provided by the authors, whereas their peers completed various handwriting exercises. The intervention spanned eight weeks with 16 sessions, each lasting 15-20 minutes. No significant differences on any measure were observed between groups from pre- to post-test. Limited effects in texting were linked to students' poor vocabulary and pinyin skills, which appeared to hinder meaningful online interaction. Consequently, many ignored the prompts, resorting to English and basic Chinese vocabulary to discuss unrelated topics. The authors conclude that typing characters should only serve to supplement CFL instruction.

Chen (2020) investigated an interactive learning platform for enhancing reading performance, adopting a single-group design. The platform included grammar videos and group-chat features, accessible on smartphones/computers. The four-week intervention involved two sessions per week: one for learning via the platform; and one for applying content through paper-based activities. Significant gains were reported in reading skills, but these findings should be viewed cautiously due to no comparator. The lack of access to the platform further impedes understanding of how instruction was precisely delivered.

Zhang (2019) also employed a single-group design to assess the potential benefits of online comics for character acquisition. Each two-hour weekly lesson involved 40-minutes of instruction via online comics from the author, with the remaining time delivered by the classroom teacher. The curriculum, based on geography content, followed a CLIL framework. The 10-week intervention included two weeks for pre-post-testing, 4-5 weeks of teaching, and 2-3 weeks for students to create their digital comics. Participants showed significant gains, with 23% achieving full marks in the post-test.

However, both Chen (2020) and Zhang (2019) lacked comparator groups, limiting the ability to attribute improvements to the instructional methods alone. Chen's (2020) study also used inconsistent testing formats by administering pre-tests online and post-tests on paper without explanation. Both assessments comprised identical test items, potentially introducing practice effects. Zhang's (2019) research lacked sufficient control over external variables, as only 40-minutes of each two-hour lesson was dedicated to the intervention, with the remainder taught by the classroom teacher. It is unclear whether linguistic gains stemmed from the intervention or were partly influenced by the classroom teacher's instruction. Harvey and Brooks (2022) indicate that novice CFL students may require more linguistic support before benefitting from typing activities. Chen (2020) and Zhang (2019) received 0*, while Harvey and Brooks (2022) obtained 2*. The above evidence suggests that texting, online comics, and interactive platforms offer no discernible advantage for CFL acquisition.

5.2.2.3 Digital writing and typing characters

One study examined digital writing, two investigated typing, and one compared both approaches. Xu et al. (2013) explored whether different CALL presentation modes enhanced orthographic knowledge, randomly assigning students to reading, animation, or digital writing conditions. Participants initially completed a reading task on individual computers, with the other two conditions counterbalanced. Each session followed a structured format: 60 characters were presented with pinyin and audio (n=20 per session); students viewed each character for 15 seconds (reading-only); they viewed stroke-by-stroke animations three times (animation group); and, in the writing condition, subjects digitally reproduced characters from memory. Pre-post-tests were conducted before and after participants completed learning all characters, followed by one-month delayed post-tests. All groups demonstrated significant gains from pre- to post-test, with the reading group significantly outperforming others in pinyin production, meaning recall, and character production, followed by animation, and digital writing. No effects were found on any condition at the delayed post-test. This study received 2* due to the absence of standardised/validated tests.

Chiu (2024) conducted a quasi-experimental crossover design comparing two instructional sequences over five weeks (eight 75-minute lessons): handwriting-to-typing and typing-to-handwriting. Each modality covered a different topic over 2.5 weeks. Both groups received identical content from the same teacher. Instruction focused on radicals and stroke-order sequencing, with handwriting exercises including copying vocabulary and constructing sentences. The typing group performed similar activities on a keyboard. While both groups improved in character recognition, the handwriting-first condition scored significantly better on both pre-post-tests, particularly in sentence construction and grammar. However, Chiu (2024) reported that 18 of 23 participants spoke L2 Chinese, suggesting heritage learner status. Although pre-test scores showed no significant differences in CFL proficiency, these findings may have limited relevance due to potential strengths in HLs' oral-aural skills. The study was rated 1*.

In a cluster RCT, Xu and Jen (2005) compared typing with pencil-and-paper handwriting to investigate the impact of a custom word-processing program on character recognition. This software differed from standard input systems by providing real-time error feedback for incorrect pinyin or missing characters in sentences. The authors reported positive outcomes associated with typing but did not specify whether effects were statistically significant.

Details on participant backgrounds and assessments were also not provided. Given that the study was conducted over three years across beginner CFL programs in America, enrolment and participant demographics likely fluctuated annually. Omitting such information limits causal interpretation, thereby receiving 0* for trustworthiness.

Tsai (2014) compared pencil-and-paper handwriting against three conditions implementing orthographic animation with audio: digital writing without feedback; typing only; and digital writing with immediate error-specific feedback. Students learned eight characters per lesson (7-9 strokes each) across 10 weekly 30-minute lessons over 16 weeks. Each group began with orthographic animation and audio, followed by weekly writing exercises completed in their assigned condition, using the eight characters taught in each session. The order of treatment conditions was counterbalanced. Only post-tests were administered after participants had completed all learning conditions, evaluating character recognition, production, and stroke-order knowledge. Findings revealed that handwriting, compared to typing-only, achieved significantly higher scores in accuracy and stroke-order awareness, while immediate error-specific feedback demonstrated no significant effects across all assessment measures. However, without pre-tests, it is difficult to discern the actual effects of each condition. The fact that students were recruited in intact classes raises concerns about whether they began the intervention at comparable levels, thereby rated 1* for trustworthiness.

Given their methodological weaknesses, the above findings offer minimal practical guidance on whether CFL teachers should prioritise digital writing, pinyin typing, or pencil-and-paper handwriting. Results are mixed, with digital writing performing comparatively worse than reading and animation (Xu et al., 2013), Chiu (2024) favouring handwriting-first, while Xu and Jen (2005) suggest that typing is more beneficial than handwriting for character learning. Indeed, the ecological validity of Xu and Jen's custom program is limited, as the software is not publicly available. Tsai (2014) further concludes that handwriting is significantly better than typing-only for stroke-order awareness. Thus, these studies provide insufficient evidence to conclude whether digital writing or typing is necessarily more superior to conventional methods like reading-only and handwriting.

5.2.2.4 Kinaesthetic-haptic learning

Xu and Ke (2020) was the only study to examine the effects of kinaesthetic-haptic learning, combining imagery with full-body movement, on orthographic knowledge. Participants completed the experiment individually in a multimedia studio equipped with a Microsoft Kinect V2, projector, and television. Students used full-body gestures to mimic character forms. The control group learned the same vocabulary by clicking through content on a laptop. While immediate post-test scores showed no significant differences between groups, the kinaesthetic-haptic group significantly outperformed their counterparts on the three-day delayed post-test, suggesting better long-term retention. However, the content was limited to characters with six or fewer strokes, as only simple character forms could be represented through body movements. It is unclear if this approach would yield similarly favourable outcomes for characters with greater stroke complexity. CFL teachers should also consider the financial feasibility of replicating the required technological setup. The study was rated 2* for trustworthiness.

5.2.2.5 Animated orthographic input and text vocalisation

Two studies investigated whether video demonstrations with etymological explanations increased orthographic knowledge. He and Huang (2014) compared animations showing the structural evolution of characters with two paper-based conditions: one included the English meaning and character only, while the other added etymological information via images. Each animation was accompanied by pictures illustrating the character's meaning (e.g. an image of an apple for the character meaning 'apple') and its English definition, without pinyin or audio. Pre-post-tests involved selecting the correct character for an English word, English-to-Chinese translations, and matching characters to pictures. The intervention lasted one 30-minute session. Pre-test scores showed no significant between-group differences. Post-tests revealed that animation significantly outperformed both paper-based conditions in character recognition. Including etymological images improved orthographic knowledge over text-only instruction, but this effect was only significant on picture-based assessments. These results, however, only apply to characters with fewer than eight strokes, while longer-term outcomes remain unknown. This study received 1* due to small sample size.

Wang (2005) conducted a cluster RCT evaluating text-with-audio, text-with-animation, and text with both audio-and-animation, against traditional instruction using printed-text only. The author utilised textbook software previously adopted in Chinese classes, which provided explanations through both words and images, and included audio narration instead of on-screen text. Therefore, the 21 compound characters students learned in the intervention were pre-selected by the software, taught in one 50-minute session. The audio-and-animation group scored the highest on immediate post-tests in character recognition, followed by text-plus-animation, printed-text (control), and text-plus-audio. Only text-with-animation and audio-and-animation were significantly correlated with post-test results. Although differences were non-significant, audio-and-animation illustrated the highest retention after one week, while printed-text showed the lowest. The remaining groups displayed equal retention. This study obtained 0* due to inadequate sampling in their cluster RCT.

Wu (2012) explored the effects of text vocalisation on reading performance, character/word recognition and recall. The intervention, spanning four weeks, required learners to complete reading activities via an online platform that displayed Chinese texts with clickable audio links, recorded by the author and native Mandarin speakers. The comparison group received text-only versions of the same material. Post-test results showed that students exposed to vocalised text significantly outperformed peers in word recognition, recall, and reading speed, though this was not significant on overall reading comprehension. The comparison group's performance remained stable throughout the intervention. Mean score analysis also suggested that lower-performing students utilising vocalised text achieved significantly better outcomes, with no added benefit for high-achievers. This study was among the most rigorous (3*), adopting random allocation and reporting the internal reliability of assessment tools.

These findings suggest that animated orthographic input (with or without audio) can enhance CFL outcomes. This approach appears particularly successful for characters with fewer than nine strokes (He & Huang, 2014), though long-term benefits remain unclear (Wang, 2005). Wu (2012) indicates that text vocalisation is more effective for lower-proficiency CFL readers. Given that all studies except Wu (2012) were rated as high RoB (0-1*), the advantages of animation compared to conventional techniques like paper-based materials should be treated cautiously. Despite positive results, the literature is simply insufficient to promote vocalised text as a superior strategy to printed-text, with only one of the 30 included studies examining this strategy.

5.2.3 Different methods of presenting character components

5.2.3.1 Grouping radicals

Four studies implemented this approach. Li (2004) investigated whether grouping radicals would improve reading comprehension and vocabulary knowledge, compared to rote memorisation. Participants were presented with five radicals, each supported by pictorial representations of its etymological development and English meaning. Students used these radicals to form 16 characters and completed paper-based tasks that involved combining radicals into compound characters. After eight weeks of instruction, students exposed to radicals scored significantly better than their counterparts across all measures. It should be noted, though, that while Li (2004) claimed that assessment tests were reliable, he did not report specific validation methods (e.g. through Cronbach's alpha). Participants were also self-selected, introducing the risk of selection bias (rated 1*).

In a quasi-experimental crossover design, Chang et al. (2014) adopted a similar pedagogical strategy, which was explored against pencil-and-paper handwriting, passive reading, and stroke-order writing, spanning four days of instruction (one lesson per day). The principal author initially taught 48 characters and facilitated comprehension through guided questions. Students then engaged in independent study on laptops: each item was presented chunk-by-chunk on-screen, with one demonstration lasting nine seconds. This procedure was repeated three consecutive times for each character in every session. Pre-, immediate-, post-, and two-week delayed post-tests assessed form recall and character production. Pre-tests showed that students were unfamiliar with the target characters before the intervention. While grouping radicals consistently outperformed other methods in the short-term, these advantages were not sustained over time. This study obtained 0* due to incomplete reporting on attrition.

Taft and Chung (1999) investigated the impact of introducing radical instruction at various stages on orthographic knowledge. Learners were presented with a set of 24 characters on three occasions. Sixteen radicals were used to create 72 single items, each appearing in three characters and occupying identical spatial positions (e.g. 口/kǒu/ is on the left in both 听/tīng/ and 叶 /yè/). These were displayed individually with English translations (e.g. 拉 /lā/ – PULL). Four conditions were compared: pupils in the (1) radicals-before group were first introduced to 16 radicals. They were told that radicals form the structural basis of characters and were given a chart linking them to the target characters. Students then completed 15-minutes of writing practice; (2) in the radicals-early condition, radicals were explained as

each character appeared, with participants using a chart to locate/identify them; (3) the radicals-late group received no instruction until just before their third exposure; and (4) the no-radicals group (control) learned characters without any orthographic information. The results significantly favoured the radicals-early group in character recall in both post- and one-week delayed post-tests, though this study received 0* due to no reporting on attrition.

Xu et al. (2014) compared the effectiveness of radical-grouping with an approach in which students learned 48 compound characters from eight separate radical groups (i.e. distributed condition). Forty-eight beginners and 40 intermediate CFL learners were divided into two intact classes: half from each proficiency-level received radical-grouping instruction, while the remaining students followed the distributed approach. Across four one-hour sessions, instruction was divided into: 20-minutes of teacher-guided vocabulary and comprehension exercises, without etymological/orthographic input; and independent study on laptops, where students viewed each character for around 30 seconds. The treatment group learned lexical items from two radical groups at a time, while their peers studied them across eight radical groups. First-year CFL pupils displayed significantly better recall and radical awareness on immediate post-test scores compared to their counterparts in the distributed group. Intermediate learners in both conditions also illustrated improvements in radical awareness, but this was not statistically significant. However, gains declined by the two-week delayed post-tests for both proficiency groups. The authors argue that intermediate learners' existing familiarity with combinational patterns and radicals, from two years of studying, may have influenced outcomes. They conclude that radical-grouping is most effective for novice CFL students. However, the authors used identical teaching materials for all students, potentially misaligning content with learners' proficiency levels. The study was rated 2* as not all assessments were validated/standardised.

Li (2004) and Chang et al. (2014) provide evidence that radical-grouping may be more efficient than conventional strategies, such as rote memorisation, handwriting, stroke-order practice, and passive reading. Taft and Chung (1999) found that teaching radicals alongside new characters improved orthographic knowledge, while Xu et al. (2014) observed that radical-grouping led to better outcomes than learning characters across multiple radical sets. However, it is unclear to what extent this approach promotes knowledge retention. High RoB in most of these studies limits the ability to offer meaningful recommendations for classroom practice or determine whether radical-grouping is indeed superior to traditional methods.

5.2.3.2 Colour-coding and flashcards

Five studies employed colour-coding techniques, one of which examined its use specifically on flashcards. In addition to investigating delayed character exposure, Osborne (2016, 2018) and Osborne et al.'s (2020, 2022) compared colour-coding with rote memorisation and a unity curriculum (control) approach (focusing on listening, speaking, reading, writing), to explore its impact on beginner learners' reading, form recognition, and recall. Lexical items were colour-coded by tone marks: neutral in pencil, and green, black, blue, and red for tones 1-4 respectively. In Osborne (2016), rote memorisation resulted in the highest recall and recognition scores at both mid- and post-tests, followed by colour-coding, the control group, and delayed exposure. Conversely, Osborne (2018) found colour-coding the most effective, yielding the highest gains in orthographic knowledge from pre- to post-test, followed by delayed instruction, rote memorisation, and the control group. The author does not clarify whether the above findings were statistically significant. In both studies, she attributes higher scores to a combination of the instructional approach and students' independent study outside class, making it difficult to isolate intervention effects.

Osborne et al. (2020) observed that different teaching strategies produced distinct linguistic outcomes. Rote memorisation achieved significantly better than other methods in character recognition at both mid- and post-tests, while colour-coding was most effective for pinyin transcription. Delayed instruction performed significantly worse across all measures in recall, recognition, and production, which the authors associate with insufficient teaching time. The control group also demonstrated limited gains, displaying a slight but non-significant increase in lexical diversity during sentence production. Similar findings were observed in Osborne et al. (2022), with results in orthographic knowledge significantly favouring rote memorisation and colour-coding from mid- to post-test, surpassing delayed exposure and the control group. Post-hoc analyses showed no significant gains in CFL outcomes for the latter two conditions. While the authors recommend combining rote memorisation and colour-coding for character teaching, they acknowledge that assessment measures primarily targeted character-centred rather than sentence-level tasks. However, the methodological strength of these experiments are weak, with Osborne et al. (2022) receiving the highest rating of 1*. Only Osborne (2018) adopted an RCT design, whereas others were quasi-experimental and lacked pre-tests, only assessing progress mid-way and post-intervention. This restricts the ability to determine instructional effectiveness due to unknown baseline proficiency. Of the approaches explored, colour-coding and rote memorisation appear more successful for CFL development.

Chung (2007) conducted two experiments evaluating the effects of different orthographic presentation formats on character acquisition. Sixteen two-character Chinese nouns were shown on flashcards under four conditions: character-pinyin-English; character-English-pinyin; pinyin-English-character; and English-pinyin-character. Participants experienced all presentation conditions in a counterbalanced order and received individual instruction from the author, attending as many sessions as needed to learn all lexical items. Results showed that placing the character to the left of the English word significantly enhanced both short- and long-term retention. Character pronunciation appeared to significantly improve when characters preceded its pinyin. Placing the English word adjacent to the character also consistently resulted in significantly better results in meaning acquisition than other formats.

In a follow-up experiment, Chung (2007) examined if colouring the pinyin or English word would increase students' orthographic awareness. Sixteen two-character words were written on flashcards, following the same teaching procedure as the first experiment. Using colour for both pinyin and English words significantly facilitated pronunciation and meaning acquisition, compared to no colouring at all. However, a major limitation in both experiments is the fact that students were permitted to attend unlimited sessions for character learning. The author does not specify how many sessions each participant completed. This potential discrepancy in treatment intensity raises concerns about whether such positive results reflect the actual efficacy of colour-coding and varied presentation formats or simply differences in instructional exposure. Both experiments were rated 0* for trustworthiness.

Given the methodological flaws outlined, there is no compelling evidence suggesting that colour-coding or rote learning is necessarily better than an integrated or delayed approach. Similarly, positioning characters before English words or after pinyin on flashcards does not show greater benefits beyond those explained by time and exposure.

5.2.3.3 Meaningful interpretation and chunking

Xu and Padilla (2013) was the only study to adopt this approach, comparing meaningful interpretation and chunking with rote repetition and stroke-order writing. Participants at beginner, intermediate, and advanced proficiency were matched by pre-test scores and randomly assigned to the treatment or control condition. Students completed a pre-session on the origins and structural types of Chinese characters. The teacher, who was also the principal author, provided examples of characters with shared radicals and illustrated how these could

be used to derive meaning. The core intervention spanned four days: the instructor introduced eight characters on Day 1, drawing on pre-session content to offer memorisation strategies; on Day 2, students learned a new set of eight characters, generating their own interpretations and practising visual chunking (i.e. recalling characters with shared radicals); Days 3 and 4 were dedicated to independent study of unfamiliar lexical items. The control group followed the same structure but focused on repetitive stroke-order writing. Post-test findings revealed that the treatment group performed significantly better than their peers across all measures on character recognition and recall, but these gains declined by the two-month delayed post-test. However, it is noteworthy that only transparent characters were taught, which may have favoured the treatment condition. These characters are inherently easier to decode since their phonetic and semantic radicals align with its pronunciation and meaning. Nonetheless, the study was methodologically robust (3*), employing an RCT design and accounting for confounding variables like heritage learner background. Indeed, no significant interactions were found between proficiency, heritage status, and treatment at either pre- or post-test, though beginner and non-heritage students consistently underperformed relative to their peers, suggesting the need for targeted support.

5.2.3.4 Etymological explanations of characters

Two studies compared teaching characters through verbal etymological explanations versus pictorial explanations. In Li and Tong (2020), lessons for the treatment group included: 5-minutes introducing the learning objectives; 25-minutes of instruction on new vocabulary with their corresponding pictograms and etymological details; and 10-minutes of review where the instructor (also the principal author) read a story featuring target characters. The comparison group received identical instruction except during the 25-minute teaching period, where characters were introduced with simple images (e.g. a picture of fruit for the character meaning ‘fruit’) instead of pictograms/etymological explanations. Students attended lessons 2-4 times per week over 19 weeks. They also engaged in 2-5 weeks of repetitive handwriting practice in-class to support independent review of target vocabulary. Despite both groups showing significant improvements in character recognition from pre- to post-test, some knowledge diminished by the one-week delayed post-test. Between-group findings illustrated that the treatment group performed significantly better at both post-tests. Only 4-5 weeks of repetition significantly boosted vocabulary recall for participants, whereas shorter intervals showed no additional benefit. Analysis of raw scores revealed that students’ ability to recall characters increased from 23% after two weeks of repetition to 53% after five weeks. The

authors conclude that, while both strategies are effective, incorporating pictograms and etymological explanations offer distinct advantages for character acquisition and retention. However, it is unclear whether positive findings resulted from the intervention, repeated practice, or a combination of both. The study received 0* for failing to report attrition.

In a single-group design, Shen (2010) conducted two experiments to compare the effects of etymological explanations versus image-based instruction on learning concrete and abstract words. Experiment 1 involved presenting 10 concrete items with etymological explanations via PowerPoint. The teaching procedure included: students repeating each word aloud three times; practising asking/answering questions in pairs; and identifying the correct flashcards after hearing the instructor's pronunciation. This was repeated with another 10 concrete words, this time using images-only for orthographic learning. This procedure was used in Experiment 2 to teach abstract items, recruiting the same pupils. Post-tests on pinyin transcriptions and form recognition were administered over two days: immediately after each 10-word set (Day 1) and again the next day (Day 2). Experiment 1 observed no significant differences between pedagogical approaches for concrete words, while Experiment 2 showed that image-only instruction significantly improved recall. The author suggests that concrete words are naturally more imageable and likely to be linked to visual representations already present in learners' mental lexicons, reducing the added benefit of pictorial support. Both experiments obtained 0* due to no comparison groups.

Evidence supporting etymological explanations over picture-based instruction remains inconclusive. Improvements in Li and Tong (2020) may be attributable to increased repetition and exposure, while Shen's (2010) findings are limited due to its single-group design. It is therefore unknown if either approach is more or less successful for CFL learning.

5.2.4 Holistic approaches

Ren (2004) adopted a cluster RCT design to compare an OVAL-writing approach with PowerPoint instruction. This alternative approach involved five steps: observing character formation; visualising its form; articulating/explaining its structure aloud; listening to the teacher's pronunciation; and practising writing while repeating the word in English and Chinese. The intervention comprised three stages: the author introduced 27 target characters; pupils completed OVAL-writing at spaced intervals (10-minutes, 24 hours, eight days, 13 days, and 15 days); and retention was assessed via written cue-dependent tasks. The control

group followed PowerPoint slides, with no spaced practice. Post-tests were administered after each interval for the treatment group, while their peers were tested only after 10-minutes of learning (T1) and post-intervention (T2=Day 15; T3=Day 16). OVAL-writing students consistently outperformed their peers at each stage (T1-T3), but no statistical analyses were conducted. The author concludes that this approach improves recall. However, prior CFL proficiency and pre-tests were not included, as well as details on the writing script and types of characters taught. With only two comparison cases in its cluster RCT, the study is severely underpowered, thereby receiving 0* for trustworthiness. Due to these shortcomings, this method appears to offer minimal value compared to PowerPoint instruction.

5.2.5 Conventional techniques

Poole and Sung (2015) was the only study to compare all conventional methods. In their quasi-experimental design, the authors evaluated the effects of pinyin-only, handwriting practice, and reading-only instruction on character recognition and oral proficiency. Students attended four 30-minute sessions, learning 50 words altogether: 20 in the first session and 10 in each subsequent session. Participants followed the same curriculum and were taught by the principal author. The pinyin-only group learned content via PowerPoint slides with pinyin displayed above each character. As the emphasis was on tonal acquisition and pronunciation, characters were not taught. The handwriting condition were instructed on stroke-order writing and radical functions. The reading-only group focused on form recognition and oral fluency, with no character production. Activities included character-pinyin matching and discussions on memorisation strategies. Instruction was delivered in Mandarin, unlike the other groups that received instruction in English, although the authors do not provide an explanation for this change. Post-tests were conducted after the third and fourth sessions. In oral proficiency, the pinyin-only group achieved the highest, followed by reading-only and handwriting. This trend was reversed for character recognition. The authors attribute these findings to differences in instructional focus/time allocated to oral and orthographic exposure across conditions. That is, given the lack of orthographic instruction, one might expect the pinyin-only group to perform comparatively worse in character recognition. This study suffers from several limitations: assessment measures were not validated; only nine pupils were recruited; no pre-tests or descriptions of prior CFL proficiency were provided; no statistical analyses were conducted; and variations in instructional focus across conditions make it challenging to draw reliable comparisons, thereby obtaining 0* for trustworthiness. The evidence does not clearly favour one approach over another for improving linguistic outcomes.

5.3 What specific methods or pedagogic approaches to teaching Chinese orthography are most effective for CFL learners at different proficiency levels?

5.3.1 Beginners

The current literature indicates substantial variation in the effectiveness of different teaching strategies for CFL beginners. Studies examining the timing of orthographic exposure suggest that delayed instruction can benefit oral proficiency and pinyin skills (Osborne, 2016, 2018; Osborne et al., 2020, 2022; Packard, 1990). In contrast, early exposure appears to provide advantages in reading (Knell & West, 2017) and character recognition (Ke & Dubravac, 2021). Harvey and Brooks (2022) additionally conclude that typing, when compared to handwriting, does not yield greater gains in reading and production skills among primary-school students with limited vocabulary knowledge. This indicates that CFL learners may need to demonstrate a certain level of proficiency before potentially observing significant improvements. Chiu's (2024) comparison of handwriting and typing offers tentative evidence that secondary pupils require more foundational orthographic support before completing typing activities, though caution is warranted in interpreting these results, as the author does not control for the potential confounding variable of heritage learners. Xu and Jen (2005) reported positive outcomes from their custom word-processing program for CFL development, but incomplete reporting and lack of access to the program restrict replicability and practical application. This finding favouring typing over handwriting appear to contradict those from Harvey and Brooks (2022) and Chiu (2024), possibly due to Xu and Jen's inclusion of adult CFL students, who may possess greater cognitive abilities than those of younger learners in the other studies. Xu and Ke (2020) also claimed that kinaesthetic-haptic learning facilitated character recognition and retention, particularly for characters with six or fewer strokes. However, its efficacy for more complex characters is unknown and the technological setup raises questions about classroom feasibility. Other CALL approaches, including animated orthographic input and pairing text with audio, seem to show greater success than handwriting and reading-only strategies in promoting both short- and long-term acquisition (He & Huang, 2014; Wang, 2005; Wu, 2012). Although digital games, interactive learning platforms, and online comics also enhanced CFL learning, these studies adopted single-group designs, limiting the ability to ascertain its benefits against other techniques (Chen, 2020; Poole et al., 2022; Zhang, 2019). Only one experiment compared digital games to PowerPoint slides, with results favouring CALL (Wang, 2024).

Findings further suggest that teaching characters through common radicals can improve outcomes (Chang et al., 2014; Li, 2004; Xu et al., 2014), especially when radicals are explained as they are presented (Taft & Chung, 1999). Although meaningful interpretation and chunking may foster temporary gains in recognition and recall, this method does not appear to increase knowledge retention, highlighting the need for greater exposure and repetition (Xu & Padilla, 2013). Li and Tong (2020) found that etymological explanations with pictorial support was better than merely introducing the character's form and English translation. These results were reaffirmed by Shen (2010), noting greater recall for abstract words but this advantage did not extend to concrete items. Colour-coding and rote learning were also associated with improved orthographic knowledge and oral skills, though long-term benefits were not assessed (Osborne, 2016, 2018; Osborne et al., 2020, 2022). Chung's (2007) experiments indicate that placing English translations on the left of characters on flashcards enhances meaning acquisition, and colour-coding pinyin/English facilitates orthographic awareness. However, it is noteworthy that inconsistent treatment intensity may have influenced outcomes, with participants attending as many sessions as required to learn the target items. Ren (2004) reported increased character retention with an OVAL-writing method compared to PowerPoint slides, but the study's small sample, with only two cases in its cluster RCT, critically hinders statistical power. Poole and Sung's (2015) investigation of traditional techniques revealed that pinyin-only instruction was most successful for speaking compared to handwriting and reading, whereas handwriting yielded the greatest outcomes in character recognition, followed by reading. The evidence presents an optimistic view of teaching approaches that appear effective for beginners, but educators should exercise caution when evaluating these findings given their methodological shortcomings.

5.3.2 Intermediate and advanced learners

Research on CFL pedagogy involving intermediate and advanced learners was scarce, with evidence drawn from only two studies. Xu et al. (2014) found no significant advantage to grouping radicals when teaching characters to second-year intermediate CFL learners (the authors did not recruit advanced students). This was compared with a distributed approach, where subjects learned across eight radical sets. The authors suggest that participants' prior exposure to radicals and combinational patterns likely influenced outcomes. Conversely, results from Xu and Padilla (2013) demonstrated that meaningful interpretation and chunking can benefit both intermediate and advanced students, leading to short-term linguistic gains, but these effects diminished two months later. Differences in outcomes may derive from the

types of characters/radicals tested: Xu et al. (2014) incorporated high-frequency radicals and identical teaching materials for both first- and second-year students, while Xu and Padilla (2013) controlled for memory load, number of strokes, character frequency, and phonetic transparency to enable more reliable within-subject comparisons. The available evidence, however, is insufficient to strongly establish whether any specific strategy is especially effective for intermediate and/or advanced learners.

Chapter 6: Conclusion

6.1 A best method for CFL pedagogy?

It is clear from the literature that various approaches to teaching Chinese orthography have been explored. The included studies evaluated a diverse range of methods, such as CALL, timing of instruction, character component presentation, and holistic approaches, often against conventional techniques like rote memorisation, handwriting, and passive reading. The findings paint an overly positive picture of the effects of different pedagogical strategies on character acquisition and reading comprehension, with only one study reporting non-significant outcomes (Harvey & Brooks, 2022). However, due to limitations previously noted, there is little reliable evidence indicating that any single approach improves CFL learning. Indeed, any approach appears better than no approach at all, raising concerns about the validity of claims made by authors of the included studies without rigorous evidence. This seems to contradict Zhang's (2024, p. 22) systematic review which, despite focusing solely on L1 English CFL learners and lacking RoB assessments, argues that several methods have "proved" successful for orthographic development. The potential advantages of each method, considered alongside its study quality, are critical factors when deciding whether to maintain existing CFL strategies in the classroom or implement those identified in this review. Strong claims about these effects have been presented in earlier chapters, but much of the evidence suffers from methodological flaws. The literature on CFL instruction, especially for advanced and intermediate learners, is also sparse and inconsistent, with a notable bias toward studies examining foundational character teaching at American universities. As such, little is known about strategies used to support orthographic learning at higher-proficiency levels or students from other geographic regions. This poses a serious concern for CFL research, as teachers seeking to develop advanced reading and character knowledge, or support students from non-Anglophone backgrounds, lack evidence-informed pedagogy.

Collectively, this systematic review has found that, while numerous pedagogical strategies have been examined in the literature to enhance CFL outcomes, robust evidence is severely lacking, particularly for younger and higher-proficiency students. Thus, further research on this topic is urgently needed. The legitimacy of claims by authors of the included studies that most approaches are superior to control groups or traditional methods, with no reported adverse effects, is questionable when such assertions are not grounded in reliable empirical research. Without transparent and rigorous evidence, it is problematic to offer confident suggestions for sound classroom practice.

6.2 Limitations

One limitation of this review was confining the search process to Western educational databases, which may have excluded relevant studies published in Chinese databases (e.g. China National Knowledge Infrastructure). It is worth noting, though, that the aim of this review was to identify strategies effective for CFL learners, who are typically based outside China, and thus may not necessarily be the focus of studies published in Chinese-language academic journals. Nonetheless, replications or future updates to this review can consider broadening the range of databases included in the search strategy to capture a wider selection of relevant literature.

While Macaro (2019) advises that systematic reviews should be conducted collaboratively, this review was predominantly completed by one individual. This inevitably introduces some degree of subjectivity and potential bias when collecting and synthesising data. Variations in personal judgement may have influenced how RoB was evaluated when interpreting Gorard's (2014) sieve. It is equally important to acknowledge the limitations of using Gorard's sieve to assess trustworthiness of studies: (1) the organisation of the sieve means that overall ratings tend to obscure the nuanced strengths of individual studies; and (2) while its intentional vagueness is designed to accommodate diverse research contexts, the lack of precise criteria (e.g. in Scale and Dropout) can complicate the rating process. Therefore, efforts were made to minimise subjectivity by recruiting another researcher to blind-screen 10% of the study abstracts, full-texts, and RoB assessments.

6.3 Implications

These findings can help CFL teachers revisit and reflect upon their current practices in the classroom, as well as be inspired to experiment with different ways of promoting character acquisition and reading performance. However, research in CFL education is still under-developed, especially in primary- and secondary- settings and among higher-proficiency learners in diverse geographic contexts. Future studies should focus on these populations within language programs that minimise selection bias, to better understand how various teaching strategies affect students who may not be motivated or influenced to learn Mandarin. More longitudinal studies are required to explore the lasting impact of different instructional approaches, particularly regarding knowledge retention. Transparency in the type of script used (i.e. simplified or traditional) is also needed to maximise replicability and ensure accurate interpretations of findings. Nonetheless, the results may prompt teachers to adjust their approaches while considering factors like resource availability and logistical constraints. Although studies rated 3* suggest benefits to grouping radicals, meaningful interpretation and chunking, and text vocalisation for improving orthographic knowledge and reading skills, the lack of high-quality research hinders strong recommendations for best practice. The strategies evaluated in this review may serve as a valuable starting point, but teachers should be wary and avoid placing undue confidence in their efficacy.

References

* Denotes publications selected as included studies in this review.

Allen, J. R. (2008). Why learning to write Chinese is a waste of time: A modest proposal. *Foreign Language Annals*, *41*(2), 237-251.

Asia Society of Australia. (2022). *Asia literacy and employability*. Retrieved from https://asiasociety.org/sites/default/files/2022-12/Asia%20Literacy%20and%20Employability.pdf

Boland, A., Cherry, R., & Dickson, R. (2014). *Doing a systematic review: A student's guide*. Sage Publications.

Boltz, W. (1994). *The origin and early development of the Chinese writing system*. American Oriental Society.

British Council. (2023). *Language trends*. Retrieved from https://www.britishcouncil.org/sites/default/files/language_trends_england_2023.pdf

British Council. (2024). *Mandarin Excellence Programme*. Retrieved from https://www.britishcouncil.org/school-resources/languages/mandarin-excellence-programme

Cao, F., Rickles, B., Vu, M., Zhu, Z., Chan, D. H. L., Harris, L. N., Stafura, J., Xu, Y., & Perfetti, C. A. (2013). Early stage visual-orthographic processes predict long-term retention of word form and meaning: A visual encoding training study. *Journal of Neurolinguistics*, *26*(4), 440-461.

Chalmers, H. (2016). *Can education learn from evidence-based medicine?*. Centre for Evidence-Based Medicine. Retrieved from https://ora.ox.ac.uk/objects/uuid:28bac2d8-0c15-4b9b-8251-e17200c7062c

Chan, J., Woore, R., Molway, L., & Mutton, T. (2022). Learning and teaching Chinese as a foreign language: A scoping review. *Review of Education*, *10*(3), 1-35.

*Chang, L. Y., Xu, Y., Perfetti, C. A., Zhang, J., & Chen, H. C. (2014). Supporting orthographic learning at the beginning stage of learning to read Chinese as a second language. *International Journal of Disability, Development and Education*, *61*(3), 288-305.

Chen, H. C. (1992). Reading comprehension in Chinese: Implications from character reading times. *Advances in Psychology*, *90*(1), 175-205.

Chen, J., Luo, R., & Liu, H. (2017). The effect of pinyin input experience on the link between semantic and phonology of Chinese character in digital writing. *Journal of Psycholinguistic Research*, *46*(4), 923-934.

*Chen, R. (2020). *The use of UMU interactive platform in Chinese language reading skills of Grade 10 Thai students*. [Master's thesis, Rangsit University]. Thailand.

*Chiu, C. (2024). *How does writing instruction impact Chinese as a foreign language learners' literacy?*. [Doctoral thesis, University of Delaware]. United States of America.

*Chung, K. H. (2007). Presentation factors in the learning of Chinese characters: The order and position of Hanyu Pinyin and English translations. *Educational Psychology*, *27*(1), 1-20.

Coltheart, M. (2005). Modelling reading: The dual-route approach. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 6-23). Blackwell.

DeFrancis, J. (1984). *The Chinese language: Fact and fantasy*. University of Hawaii Press.

Department of Continuing Education. (n.d.). *Chinese language courses*. University of Oxford. Retrieved from https://www.conted.ox.ac.uk/about/chinese-language-courses

Dos Santos, L. M. (2024). Motivations for learning Chinese as a foreign language: A case study in Belgium. *International Journal of Instruction*, *17*(2), 85-104.

Duff, P.A., Anderson, T., Ilnyckyj, R., Van Gaya, E., Wang, R., & Yates, E. (2013). *Learning Chinese: linguistic, sociocultural, and narrative perspectives*. De Gruyter.

Duff, P. A., & Li, D. (2004). Issues in Mandarin language instruction: Theory, research, and practice. *System*, *32*(1), 443-456.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 63-103). Blackwell.

Erbaugh, M. S. (2022). The acquisition of Mandarin. In D. I. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition* (pp. 373-455). Psychology Press.

Ésik, S. (2020). Teaching Chinese characters to second language learners. *Researching and Teaching Chinese as a Foreign Language*, *3*(1), 1-22.

Everson, M. E. (1988). Speed and comprehension in reading Chinese: Romanisation vs. characters revisited. *Journal of the Chinese Language Teachers Association*, *23*(1), 1-15.

Everson, M. E. (1994). Toward a process view of teaching reading in the second language Chinese curriculum. *Theory into Practice*, *33*(1), 4-9.

Everson, M. E. (2011). Best practices in teaching logographic and non-roman writing systems to L2 learners. *Annual Review of Applied Linguistics*, *31*(1), 249-274.

Everson, M. E., & Xiao, Y. (2009). *Teaching Chinese as a foreign language: Theories and applications*. Cheng & Tsui Company.

Gil, J. (2017). *Soft power and the worldwide promotion of Chinese language learning: The Confucius Institute project*. Multilingual Matters.

Gil, J. (2024). Confucius Institute and Confucius Classroom closures: Trends, explanations, and future directions. *Applied Linguistics Review*, *15*(2), 699-712.

Goh, Y. S. (2017). The spread of Mandarin as a global language. In Y. S. Goh & Y. Wu (Eds.), *Teaching Chinese as an international language: A Singapore perspective*. Cambridge University Press.

Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics*, *110*, 47-60. http://www.radstats.org.uk/no110/Gorard110.pdf

Gorard, S. (2024). Judging the relative trustworthiness of research results: How to do it and why it matters. *Review of Education*, *12*(1), 1-14.

Gough, D., Thomas, J., & Oliver, S. (2017). *An introduction to systematic reviews*. Sage Publications.

Han, Z. (2016). Research meets practice: holding off and holding on. *Chinese as a Second Language, 51*(3), 236-251.

Hao, P., & Li, F. (2024). Exploring the challenges of learning and teaching Chinese/Mandarin language at higher education institutes: Voices from non-Chinese speaker teachers and learners. *Journal of Psycholinguistic Research*, *53*(6), 1-17.

Hartig, F. (2012). Confucius Institutes and the rise of China. *Journal of Chinese Political Science*, *17*(1), 53-76.

*Harvey, R. E., & Brooks, P. J. (2022). Effects of text messaging using digital Pinyin input on literacy skills of elementary Chinese immersion learners. *Language Teaching Research*, *0*(0), 1-27.

He, C. (2023). Mediation effects of character-related tasks between instruction type and character writing in delayed character introduction for Chinese as a foreign language learners. *Journal of Education and Educational Research*, *2*(2), 100-103.

*He, J., & Huang, H. (2014). Learning Chinese characters with animated etymology. *International Journal of Computer-Assisted Language Learning and Teaching*, *38*(1), 99-118.

Honorof, D., & Feldman, L. (2006). The Chinese character in psycholinguistic research: Form, structure, and the reader. In P. Li, L. H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *The handbook of East Asian psycholinguistics* (pp. 195-217). Cambridge University Press.

Hsiung, H. Y., Chang, Y. L., Chen, H. C., & Sung, Y. T. (2017). Effect of stroke-order learning and handwriting exercises on recognising and writing Chinese characters by Chinese as a foreign language learners. *Computers in Human Behaviour*, *74*(1), 303-310.

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

Huang, S. (2022). A tale of two less successful CSL readers. In L. Li & D. Zhang (Eds.), *Reading in Chinese as an additional language* (pp. 156-178). Routledge.

Kan, Q., Owen, N., & Bax, S. (2018). Researching mobile-assisted Chinese-character learning strategies among adult distance learners. *Innovation in Language Learning and Teaching*, *12*(1), 56-71.

*Ke, S., & Dubravac, S. (2021). When should characters be introduced to novice-level Chinese in a blended learning setting?. *Studies in Chinese Learning and Teaching*, *6*(1), 1-26.

*Knell, E., & West, H. (2017). To delay or not to delay: The timing of Chinese character instruction for secondary learners. *Foreign Language Annals*, *50*(3), 519-532.

Kwoh, S. (2007). Mainstreaming and professionalising Chinese-language education: A new mission for a new century. *Chinese America: History and Perspectives*, *1*(1), 261-264.

Lam, H. (2011). A critical analysis of the various ways of teaching Chinese characters. *Electronic Journal of Foreign Language Teaching*, *24*(4), 496-518.

*Li, J. T., & Tong, F. (2020). The effect of cognitive vocabulary learning approaches on Chinese learners' compound word attainment, retention, and learning motivation. *Language Teaching Research*, *24*(6), 834-854.

Li, M. (2020). A systematic review of the research on Chinese character teaching and learning. *Frontiers of Education in China*, *15*(1), 39-72.

*Li, W. (2004). *The "grapheme combination method": Teaching and learning Chinese characters through associative links*. [Doctoral thesis, George Mason University]. United States of America.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Annals of Internal Medicine*, *151*(4), 65-94.

Lin, Y. H. (2007). *The sounds of Chinese*. Cambridge University Press.

Liu, G. Q., & Lo Bianco, J. (2007). Teaching Chinese, teaching in Chinese, and teaching the Chinese. *Language Policy*, *6*(1), 95-117.

Lü, X., Ostrow, K. S., & Heffernan, N. T. (2019). Save your strokes: Chinese handwriting practice makes for ineffective use of instructional time in second language classrooms. *AERA Open*, *5*(4), 1-15.

Lü, X., & Zhang, J. (1999). Reading efficiency: A comparative study of English and Chinese orthographies. *Literacy Research and Instruction*, *38*(4), 301-317.

Ma, X., Gong, Y., Gao, X., & Xiang, Y. (2017). The teaching of Chinese as a second or foreign language: A systematic review of the literature 2005-2015. *Journal of Multilingual and Multicultural Development*, *38*(9), 815-830.

Macaro, E. (2019). Systematic reviews in applied linguistics. In J. Mckinley, & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 230-239). Routledge.

Mayumi, K., & Zheng, Y. (2023). Becoming a speaker of multiple languages: An investigation into UK university students' motivation for learning Chinese. *The Language Learning Journal*, *51*(2), 238-252.

McBride, C. A. (2015). Is Chinese special? Four aspects of Chinese literacy acquisition that might distinguish learning Chinese from learning alphabetic orthographies. *Educational Psychology Review*, *28*(3), 523-549.

McGinnis, S. (1999). Students' goals and approaches. In M. Chu (Ed.), *Chinese language teachers' association monograph series: Mapping the course of the Chinese language field* (pp. 151-168). Chinese Language Teachers.

McHugh, M. L. (2012). Interrater reliability: The Kappa statistic. *Biochemia Medica*, *22*(3), 276-282.

Moher, D., Pham, B., Lawson, M., & Klassen, T. (2003). The inclusion of reports of randomised trials published in languages other than English in systematic reviews. *Health Technology Assessment*, *7*(41), 1-98. https://doi.org/10.3310/hta7410

Mori, Y. (1998). Effects of first language and phonological accessibility on kanji recognition. *The Modern Language Journal*, *82*(1), 69-82.

Murphy, V. (2014). *Second language learning in the early school years*. Oxford University Press.

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Newman, M., & Gough, D. (2020). Systematic reviews in educational research: Methodology, perspectives, and application. *Systematic Reviews in Educational Research*, *64*(3), 3-22.

Nicoletti, R., & Culligan, K. (2022). *The Mandarin Excellence Programme: Evaluation of the first five years*. Retrieved from https://ci.ioe.ac.uk/wp-content/uploads/2022/05/MEP-Independent-Evaluation-Report-2016-2021.pdf

Norman, J. (1988). *Chinese*. New York: Cambridge University Press.

NSW Education Standards Authority. (n.d.). *Modern languages K-10 syllabus*. Retrieved from https://curriculum.nsw.edu.au/learning-areas/languages/modern-languages-k-10-2022/overview

Oakley, A. (2007). In M. Hammersley (Ed.), *Educational research and evidence-based practice* (pp. 91-105). Sage Publications.

Orton, J. (2011). Educating Chinese language teachers – some fundamentals. In L. Tsung & K. Cruickshank (Eds.), *Teaching and learning Chinese in global contexts: Multimodality and literacy in the new media age* (pp. 151-164). Continuum.

Orton, J. (2016). *Building Chinese language capacity in Australia*. Australia China Research Institute. Retrieved from https://www.australiachinarelations.org/sites/default/files/20032%20ACRI%20Jane%20Orton%20-%20Chinese%20Language%20Capacity_web_0.pdf

*Osborne, C. (2016). Chinese in the classroom: Initial findings of the effects of four teaching methods on beginner learners. *Journal of Second Language Teaching and Research*, *5*(1), 202-225.

*Osborne, C. (2018). Examining character recognition and recall skills of CFL beginner learners under four different approaches. *Journal of the Irish Association for Applied Linguistics*, *25*(1), 52-73.

*Osborne, C., Zhang, Q., & Adamson, B. (2022). The next steps for teaching characters in CFL: Investigating the effects of four Character teaching methods on beginner learners. *International Journal of Chinese Language Education*, *11*(1), 45-82.

*Osborne, C., Zhang, Q., & Zhang, G. X. (2020). Which is more effective in introducing Chinese characters? An investigative study of four methods used to teach CFL beginners. *Language Learning Journal*, *48*(4), 385-401.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan – A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 1-10. https://doi.org/10.1186/s13643-016-0384-4

*Packard, J. L. (1990). Effects of time lag in the introduction of characters into the Chinese language curriculum. *Modern Language Journal*, *74*(1), 167-175.

Packard, J. L., Chen, X., Li, W., Wu, X., Gaffney, J. S., Li, H., & Anderson, R. C. (2006). Explicit instruction in orthographic structure and word morphology helps Chinese children learn to write characters. *Reading and Writing*, *19*(1), 457-487.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hossmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. Q., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Research Methods and Reporting*, *372*(71), 1-9.

Perfetti, C. A., & Liu, Y. (2005). Orthography to phonology and meaning: Comparisons across and within writing systems. *Reading and Writing*, *18*(3), 193-210.

Perfetti, C. A., Zhang, S., & Berent, I. (1992). Reading in English and Chinese: Evidence for a "universal" phonological principle. *Advances in Psychology*, *94*(1), 227-248.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell.

*Poole, F., Clarke-Midura, J., & Ji, S. (2022). Exploring the affordances and effectiveness of a digital game in the Chinese dual language immersion classroom. *Journal of Technology and Chinese Language Teaching*, *13*(1), 46-73.

*Poole, F., & Sung, K. (2015). Three approaches to beginning Chinese instruction and their effects on oral development and character recognition. *Eurasian Journal of Applied Linguistics*, *1*(1), 59-75.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K. & Duffy, S., (2006). Guidance on the conduct of narrative synthesis in systematic reviews. *A product from the ESRC methods programme Version*, *1*(1), p.b92.

Ramezanzadeh, A. M., & Woore, R. (2023). *Scoping Review of Teaching and Learning Arabic as an L2-Database*. Retrieved from https://ora.ox.ac.uk/objects/uuid:322b75c2-9a5b-4214-9df7-a81b1d5e848b

*Ren, G. (2004). Introducing OVAL writing: A new approach to Chinese character retention for secondary non-Chinese-speaking background learners. *In Australian Federation of Modern Language Teachers Associations National Conference*. https://search.informit.org/doi/epdf/10.3316/ielapa.200409296

Shen, H. H. (2005). An investigation of Chinese-character learning strategies among non-native speakers of Chinese. *System*, *33*(1), 49-68.

Shen, H. H. (2008). An analysis of word decision strategies among learners of Chinese. *Foreign Language Annals*, *41*(3), 501-524.

*Shen, H. H. (2010). Imagery and verbal coding approaches in Chinese vocabulary instruction. *Language Teaching Research*, *14*(4), 485-499.

Shen, H. H., & Ke, C. (2007). Radical awareness and word acquisition among non-native learners of Chinese. *The Modern Language Journal*, *91*(1), 97-111.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5-11.

Snow, D. (2013). Revisiting Ferguson's defining cases of diglossia. *Journal of Multilingual and Multicultural Development*, *34*(1), 61-76.

Starr, D. (2009). Chinese language education in Europe: The Confucius Institutes. *European Journal of Education*, *44*(1), 65-82.

Swihart, D. W. (2004). *Success with Chinese, level 1: Listening and speaking*. Cheng & Tsui Company.

*Taft, M., & Chung, K. (1999). Using radicals in teaching Chinese characters to second language learners. *Psychologia*, *42*(4), 243-251.

Tan, L. H., Spinks, J. A., Eden, G. F., Perfetti, C. A., & Siok, W. T. (2005). Reading depends on writing in Chinese. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(24), 8781-8785.

Taylor, I., & Taylor, M. M. (2014). *Writing and literacy in Chinese, Korean, and Japanese*. John Benjamins.

Tinsley, T., & Board, K. (2013). *Language learning in primary and secondary schools in England: Findings from the 2012 Language Trends survey*. CfBT Education Trust.

Tong, X., & Yip, J. H. Y. (2015). Cracking the Chinese character: Radical sensitivity in learners of Chinese as a foreign language and its relationship to Chinese word reading. *Reading and Writing*, *28*(2), 159-181.

*Tsai, C. H. (2014). *Multimedia mediation and Chinese orthographic character learning among non-heritage CFL beginners*. [Doctoral thesis, University of Iowa]. United States of America.

US Department of State (n.d.). *Foreign language training: Foreign service institute*. Retrieved from https://www.state.gov/foreign-language-training/

Valdés, G. (2001). Heritage language students: Profiles and possibilities. In J. Peyton, D. Ranard & Q.S. McGinnis (Eds.), *Heritage languages in America: Preserving a national resource* (pp. 37-77). Centre for Applied Linguistics and Delta Systems.

Walker, G. (1989). Intensive Chinese curriculum: The EASLI model. *Journal of the Chinese Language Teachers Association*, *24*(2), 43-84.

Walker, J., & Poole, F. (2016). Effects of delaying character instruction in a Chinese as a foreign language classroom on affective outcomes. *Researching and Teaching Chinese as a Foreign Language*, *2*(2), 162-180.

*Wang, L. (2005). *The impact of multimedia on Chinese learners' recognition of characters: A quantitative and qualitative study*. [Doctoral thesis, Purdue University]. United states of America.

Wang, L., & Higgins, L. T. (2008). Mandarin teaching in the UK in 2007: A brief report of teachers' and learners' views. *Language Learning Journal*, *36*(1), 91-96.

Wang, M. (2025). Teaching Chinese characters: The challenges and strategies of addressing the unique aspects of Chinese characters and exploring effective pedagogical approaches. *International Journal of Sociologies and Anthropologies Science Reviews*, *5*(2), 827-836.

*Wang, T. (2024). Designing a digital game for Chinese character learning: A theory-driven practice approach. *Education Sciences*, *14*(12), 1-14.

Wong, K. S. R. (2013). Learning to read Chinese: The importance of reading skills in a non-alphabetic language. *Perspectives on Language and Literacy*, *39*(1), 33-39.

Wong, Y. K. (2017). Relationships between reading comprehension and its components in young Chinese as a second language learners. *Reading and Writing*, *30*(5), 969-988.

*Wu, Y. (2012). Using external text vocalisation to enhance reading development among beginning level Chinese learners. *Journal of the Chinese Language Teachers Association*, *47*(1), 1-23.

Xing, J. (2006). *Teaching and learning Chinese as a foreign language: A pedagogical grammar*. Hong Kong University Press.

Xu, H. L., & Moloney, R. (2019). Motivation for learning Chinese in the Australian context: A research focus on tertiary students. In M. Lamb, K. Csizér, A. Henry, & S. Ryan (Eds.), *The Palgrave handbook of motivation for language learning* (pp. 449-469). Springer.

*Xu, P., & Jen, T. (2005). "Penless" Chinese language learning: A computer-assisted approach. *Journal of the Chinese Language Teachers Association*, *40*(2), 25-42.

Xu, W. (2023). Rising China and rising Chinese: An investigation into African international students' language beliefs. *System*, *115*(1), 1-9.

Xu, W. (2024). Pedagogic affect and African international students' attunement to Chinese language learning. *Journal of Multilingual and Multicultural Development*, *45*(7), 2369-2381.

*Xu, X., & Ke, F. (2020). Embodied interaction: Learning Chinese characters through body movements. *Language Learning & Technology*, *24*(3), 136-159.

*Xu, X., & Padilla, A. M. (2013). Using meaningful interpretation and chunking to enhance memory: The case of Chinese character learning. *Foreign Language Annals*, *46*(3), 402-422.

*Xu, Y., Chang, L. Y., & Perfetti, C. A. (2014). The effect of radical-based grouping in character learning in Chinese as a foreign language. *The Modern Language Journal*, *98*(3), 773-793.

*Xu, Y., Chang, L. Y., Zhang, J., & Perfetti, C. A. (2013). Reading, writing, and animation in character learning in Chinese as a Foreign Language. *Foreign Language Annals*, *46*(3), 423-444.

Yang, C. (2016). *The acquisition of L2 Mandarin prosody*. John Benjamins.

Yang, J. (2022). Teenage beginners' perceptions of learning Chinese characters: A case study. *Journal of Chinese Writing Systems*, *6*(11), 3-15.

Ye, L. (2013). Shall we delay teaching characters in teaching Chinese as a foreign language?. *Foreign Language Annals*, *46*(1), 610-627.

Ye, Y., & McBride, C. (2022). A dynamic interactive model of Chinese spelling development. *Educational Psychology Review*, *34*(4), 2897-2917.

Yue, Y. (2017). Teaching Chinese in K-12 schools in the United States: What are the challenges?. *Foreign Language Annals*, *50*(3), 601-620.

Zeichner, K., & Liu, K. Y. (2010). A critical analysis of reflection as a goal for teacher education. *Educação & Sociedade*, *29*(103), 535-554.

Zhang, J., Li, H., Dong, Q., Xu, J., & Sholar, E. (2016). Implicit use of radicals in learning characters for non-native learners of Chinese. *Applied Psycholinguistics*, *37*(3), 507-527.

Zhang, P. N. (2021). Typing to replace handwriting: Effectiveness of the typing-primary approach for L2 Chinese beginners. *Journal of Technology and Chinese Language Teaching*, *12*(2), 1-28.

Zhang, S. (2023). Effective character teaching methods for L1 English Chinese as a foreign language learners: A review of empirical research. *Chinese as a Second Language*, *58*(3), 205-236.

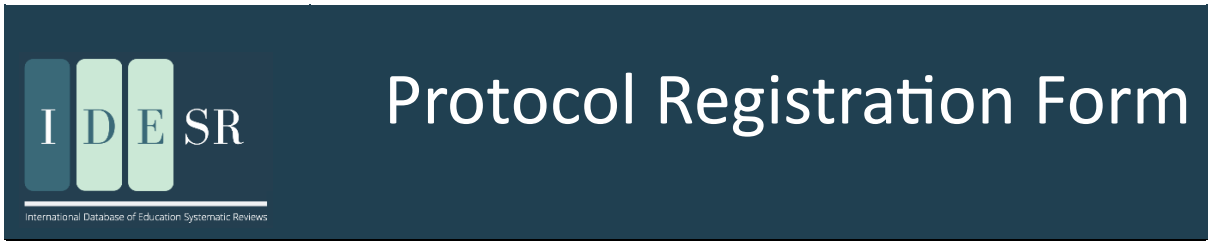
Zhang, S. (2024). Effective character teaching methods for L1 English Chinese as a foreign language learners: A review of empirical research. *Chinese as a Second Language*, *58*(2), 1-31.

*Zhang, T. (2019). *Online comics for the teaching and learning of Chinese language in the Australian context*. [Master's thesis, Western Sydney University]. Australia.

Zhang, T., & Ke, C. (2018). Research on L2 Chinese character acquisition. In C. Ke (Ed.), *The Routledge handbook of Chinese second language acquisition* (pp. 103-133). Routledge.

Zhao, Y. (2011). Review article: A tree in the wood: A review of research on L2 Chinese acquisition. *Second Language Research*, *27*(4), 559-572.

Appendix A: Protocol registration form



The content of this form is based on Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. (2015). Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1). doi: 10.1186/2046-4053-4-1. We strongly recommended that this form is completed in accordance with the recommendations in that paper.

Title

Investigating the substantive linguistic effects of alternative teaching methods on CFL learners' character acquisition and reading outcomes. Protocol for a systematic review.
The title should identify the report as a protocol of a systematic review. If this is an update of an existing review, indicate this in the title. Examples: The effectiveness of task-based language teaching for adult learners of foreign or second languages. Protocol for a systematic review. *Or* The effectiveness of written corrective feedback for the acquisition of L2 grammar among primary-aged language learners. A protocol for an update of the systematic review by Bloggs and Smith (2020).

Main Contact/Corresponding Author

Name: [redacted]
Institutional affiliation: University of Oxford
e-mail address: [redacted]
Physical mailing address: Department of Education, 15 Northam Gardens, Oxford, OX2 6PY

Additional Authors

Name: [redacted]
Institutional affiliation: University of Oxford
e-mail address: [redacted]
On the IDESR website, you will be able to add as many additional authors as you need.

Review Question(s)

1. What instructional methods or pedagogical approaches have been evaluated for effectiveness on character acquisition and reading outcomes among school- and university-aged CFL learners?
2. What are the substantive linguistic effects of these methods on character acquisition and reading outcomes?

3. What methods or approaches are most effective for CFL learners at different proficiency levels?

Provide the review questions. For reviews of interventions, include reference to PI(E)CO, as appropriate (Participants, Intervention (Exposure), Comparator, Outcomes). Examples: What are the effects of study abroad compared to classroom teaching on vocabulary acquisition among adolescent learners of a foreign language? *Or* What is the impact on academic attainment of attending a bilingual school compared to a target language only school among minority language users?

Rationale

Teaching and learning Chinese as a foreign language (CFL) has gained popularity over recent years. However, there has been limited focus on offering a transparent and comprehensive overview of the existing literature, and critical appraisals of the evidence's trustworthiness are generally lacking. This makes it difficult to provide practical guidance for relevant stakeholders seeking to improve CFL education. This review evaluates the current state of knowledge, which is particularly prevalent given the challenges involved in learning the Chinese writing system. Efforts to improve the quality of Mandarin teaching and learning have also increased in certain regions, including the United Kingdom (British Council, 2024), Africa (Xu, 2023), and Ireland (Osborne et al., 2022). Identifying effective strategies to enhance literacy outcomes among CFL learners is valuable for both researchers and teachers, as it can inform future-related investigations and strengthen the connection between research and practice. Without an objective evaluation and quality assessment of existing evidence, the efficacy of CFL teaching methods remains unclear, and teachers risk basing their pedagogical decisions on intuition and experience alone.

British Council. (2024). *Mandarin Excellence Programme*. Retrieved from https://www.britishcouncil.org/school-resources/languages/mandarin-excellence-programme

Osborne, C., Zhang, Q., & Adamson, B. (2022). The next steps for teaching characters in CFL: Investigating the effects of four character-reaching methods on beginner learners. *International Journal of Chinese Language Education*, *11*(1), 45-82.

Xu, W. (2023). Rising China and rising Chinese: An investigation into African international students' language beliefs. *System*, *115*(1), 1-9.

In no more than 300 words, describe the rationale for the review in the context of what is already known.

Inclusion Criteria

Bibliographic information

Include 1: Studies with a complete reference.

Exclude 1: Studies with insufficient or incomplete references.

Rationale: Complete bibliographic information is necessary for retrieval of full reports.

Date of publication

Include 2: No restrictions on publication date.

Rationale: Collecting as much relevant data as possible, regardless of when it was conducted.

Participants
Include 3: Studies that target typically developing foreign language learners. Include studies without explicit mention of learning ability, if it is reasonable to assume that participants are typically developing individuals.
Exclude 3: Studies that focus solely on non-typically developing learners (e.g. learners with developmental language disorders and/or identified fine motor issues).
Rationale: The review seeks to evaluate the effectiveness of teaching methods in typically developing students. Findings focusing on non-typically developing learners may lack applicability to a larger population.

Include 4: Studies conducted in primary, secondary, and university educational settings (students aged 4-5 onwards).
Exclude 4: Studies conducted in informal settings (e.g. parents teaching their children at home); learners not enrolled in formal language courses (e.g. those learning for pleasure or exchange students).
Rationale: The study focuses solely on CFL learners' linguistic outcomes in formal institutional settings. Therefore, studies conducted in other environments or involving learners not enrolled in language programs are irrelevant.

Include 5: Studies involving CFL or L2 Mandarin learners. Include studies without explicit reference to the variety, if it is reasonable to assume that students are studying the Chinese writing system (whether simplified or traditional) with Mandarin pronunciation and pinyin.
Exclude 5: Studies involving heritage learners or students in Chinese-speaking environments (e.g. Singapore), as well as those focused on varieties other than Mandarin (e.g. Cantonese).
Rationale: The review focuses on the Chinese writing system with Mandarin pronunciation and pinyin. Studies involving heritage learners, students in Chinese-speaking contexts, or learners of non-Mandarin varieties differ considerably in their prior experiences, lexis, and tonal systems, thus these results will be excluded.

Intervention
Include 6: Studies in which learners receive some form of instruction. Include studies even if part of the instruction (or intervention) requires learners to work independently (e.g. homework, rote memorisation).
Exclude 6: Studies without an instructional period and/or focus on a teaching method (e.g. case study, students are solely given independent time to learn characters).
Rationale: The paper seeks to evaluate the impact of teaching approaches; thus, the intervention must involve some form of instruction, not purely a learning intervention (e.g. independent study time, homework).

Outcomes
Include 7: Primary research studies reporting any measure of reading outcomes, including but not limited to character/word recognition, reading comprehension, orthographic and/or phonological awareness, reading fluency, meaning translations, radical/character knowledge development, or pinyin transcriptions. Include studies reporting quantitative outcomes.

Exclude 7: Systematic reviews or studies that only provide narrative analysis of an intervention without assessment measures on reading outcomes, or studies that examine non-linguistic aspects (e.g. motivation, attitudes).
Rationale: As the study is interested in examining the effects of teaching approaches on linguistic outcomes, only quantitative papers will be included.

Publication date
Include 8: Include grey literature (e.g. theses/dissertations and conference abstracts).
Exclude 8: Do not exclude studies based on publication type.
Rationale: The paper includes grey literature to minimise potential publication bias.

Study design
Include 9: Studies that attempt to identify a causal relationship (e.g. randomised controlled trials and quasi-experimental studies).
Exclude 9: Non-intervention research (e.g. case study, ethnography).
Rationale: Studies that explore causality are likely to retrieve relevant papers that assess the effectiveness of a specific intervention.

Language of publication
Include 10: Publications in any language.
Exclude 10: Do not exclude studies based on language of publication.
Rationale: Exclusion of papers in any language neglects an important body of literature that may help answer the research questions.

Moher, D., Pham, B., Lawson, M., & Klassen, T. (2003). The inclusion of reports of randomised trials published in languages other than English in systematic reviews. *Health Technology Assessment*, *7*(41), 1-98. https://doi.org/10.3310/hta7410

List here the criteria to be used for inclusion in the review. As appropriate, include information about the population, interventions, comparators, primary outcomes, setting, study design(s), time frame, publication types, language(s) of publication, etc.

Information Sources

The search for full reports will include databases from various disciplines to ensure a comprehensive review. Databases will be accessed via Oxford University's Bodleian Library subscription. Additional sources from the reference lists of previous systematic reviews will be used to supplement the search for relevant studies. This will help provide a more thorough overview of the literature and reduce the risk of neglecting important studies that bibliographic databases may miss (Liberati et al., 2009).

Databases:

- ProQuest Social Science Premium Collection (including ERIC)
- British Education Index
- Linguistics Collection (including LLBA)
- PsychINFO
- Web of Science
- SCOPUS
- ProQuest Dissertations & Theses Global

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Annals of Internal Medicine*, *151*(4), 65-94.

Describe all intended information sources. Include the names of electronic databases, journals or websites that will be hand searched, contact with study authors, grey literature sources, etc.

Search Strategy

My dissertation supervisor and two librarians at Oxford University's Department of Education were consulted to develop the initial search and Boolean strings for use in the databases mentioned above. Piloting the search strategies in Web of Science found that many results focused on Chinese learners of English as an additional language. To address this issue and refine the results, broad terms such as 'foreign language', 'Chinese language' and 'Chinese' were removed from the search strings. Terms related to participants' age, including 'primary', 'secondary' and 'tertiary', were initially tested but also later excluded, as the study focuses on all CFL learners enrolled in formal institutions. Consequently, two fields of search terms were created to represent the concepts of CFL and Chinese teaching/learning. Terms within each field were connected using the 'OR' function and each field was combined using 'AND'. The complete Boolean string comprises: ab("Chinese as a foreign language" OR "Mandarin as a foreign language" OR "Chinese character*" OR "Chinese orthograph*" OR "Chinese writing system" OR "Chinese learning") AND ab("Chinese reading" OR "Chinese writing" OR "radical knowledge" OR "Chinese literacy" OR "character recognition" OR "character recall" "character acquisition" OR "character knowledge" OR "character retention" OR "character improvement" OR reading development OR "character recall"). Minor adjustments in wording may be made depending on the database.

Present the search strategy to be used for at least one electronic database such that it could be repeated by a third party. Include planned limiters, for example, date range, and location in the text (e.g. Title, Abstract, or Full Text). Present these as a Boolean phrase if possible. If Boolean phrasing is inappropriate for your review, present the search strategy in a way that can allow replication by a third party.

Data Management

Data from the search strategy will be uploaded to Rayyan, an online management tool that allows multiple authors to compare and screen studies based on the eligibility criteria (Ouzzani et al., 2016). Bibliographic information will be organised in a word document, with notes detailing the reasons for including/excluding studies, as well as study characteristics, and relevant findings following the PICO structure (participants, intervention, comparator, outcomes) and data extraction form (see data items).

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan – A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 1-10. https://doi.org/10.1186/s13643-016-0384-4

Describe the mechanisms by which the data will be managed throughout the review. For example, say which data management software the review team use, e.g. Rayyan, EPPI Reviewer, Excel, Covidence, etc. Describe if these will change with different phases of the review (abstract screening, full text screening, data extraction, etc.).

Selection Process

Duplicates will be removed after an initial search on Rayyan. The remaining abstracts will be screened against the eligibility criteria, ensuring they meet the inclusion criteria. Abstracts lacking sufficient information will be considered for full-text screening, along with those that clearly meet the eligibility criteria. Screening will cease for abstracts that breach any one of the inclusion criteria, and the reviewer will document the reasoning for this decision using 'Exclude 1-9' labels according to the eligibility criteria. If full reports are ambiguous and prevent determining whether they meet all inclusion criteria, the principal author of the reports will be contacted for clarification.

A second reviewer studying a master's degree in the field of applied linguistics has been recruited and briefed on the study's aim and criteria. This person will review a random 10% sample of abstracts. Their screening decisions will remain concealed from the main reviewer until both have completed the process. The extent of agreement will be calculated using Cohen's Kappa. A Kappa value between 0.81-1.00 (McHugh, 2012) and an inter-rater agreement of 90% or above are generally considered high in social science research (Ramezanzadeh & Woore, 2023). If agreement falls below either threshold, both screeners will randomly review an additional 10% of abstracts and discuss any conflicts to reach a consensus. Once a Kappa value of 0.81 or higher has been reached, the remaining abstracts will be screened by the first author.

McHugh, M. L. (2012). Interrater reliability: The Kappa statistic. *Biochemia Medica*, *22*(3), 276-282.

Ramezanzadeh, A. M., & Woore, R. (2023). Scoping Review of Teaching and Learning Arabic as an L2-Database. https://ora.ox.ac.uk/objects/uuid:322b75c2-9a5b-4214-9df7-a81b1d5e848b

Describe the method by which studies will be selected for inclusion at each stage of the review. For example, how many reviewers will screen abstracts/full texts? What quality assurance procedures will be in place in each of these phases (dual screening of all records, percentage dual screened then checked for consistency, etc.)?

Data Collection Process

Data will be extracted using an adapted version from Chan et al.'s (2022) scoping review of CFL teaching and learning, as well as Ramezanzadeh and Woore's (2023) scoping review of teaching and learning L2 Arabic. The following section outlines the relevant data items and includes additional variables, such as teaching approach(es). The principal author of the reports will be contacted for further information if data is missing. A word document will be used to extract and compile relevant data from each publication.

Chan, J., Woore, R., Molway, L., & Mutton, T. (2022). Learning and teaching Chinese as a foreign language: A scoping review. *Review of Education*, *10*(3), 1-35. https://doi.org/10.1002/rev3.3370

Ramezanzadeh, A. M., & Woore, R. (2023). Scoping Review of Teaching and Learning Arabic as an L2-Database. https://ora.ox.ac.uk/objects/uuid:322b75c2-9a5b-4214-9df7-a81b1d5e848b

Describe how data will be extracted from reports. Will a data extraction form be used? Will this be piloted? Will data be extracted independently by multiple reviewers? What is the process for obtaining data not contained in the reports (e.g. contacting authors directly)?

Data Items

Administrative:

- What date was the form completed?
- Complete reference
- Publication type
- Language
- Database source
- Source of funding

Contextual:

- When was the study conducted/
- Study/research location
- Research question(s)
- Educational level
- Type of learning (e.g. intensive language program, part of school/tertiary curriculum)
- Type of institution (e.g. private school)
- Delivery (e.g. online, in-person)
- Writing script (e.g. traditional or simplified)
- Participant demographics (e.g. age, gender, L1 background, CFL proficiency, socioeconomic background)

Intervention:

- Study/research design (e.g. quasi-experimental, randomised controlled trial)
- Study duration
- Number and sample size of comparison groups (including dropout, descriptions of treatment and control groups)
- Method of participant allocation (e.g. random, intact classes, prior scores, unknown/not reported)
- Measurement tools (e.g. standardised tests, descriptions and number of tests, were they piloted, validated, and/or blind-marked?)
- Teaching approach(es) (descriptions of the lesson(s), including the instructors involved and their experiences or background, duration of each lesson, was it compared to other pedagogical approaches?)
- Teaching/learning materials (e.g. vocabulary lists, flashcards, worksheets)

Outcomes:

- What language skills/types of reading outcomes were assessed? (e.g. character knowledge, retention)
- How was data analysed? (e.g. names of statistical software)
- Effect size according to Cohen's D or Hodges' G

- Findings (e.g. mean scores, standard deviations, outcomes of pre-post intervention)
- Author(s) conclusions

List and define all data items that will be extracted (e.g. participant info, outcome measures, sources of funding, study design, etc.).

Risk of bias/trustworthiness of individual studies

Gorard's (2014) 'sieve' will be used to evaluate the methodological quality of each study. This framework supports academics in assessing the validity of outcomes in social science intervention research using a star rating system, ranging from zero (indicating a study with high risk of bias) to four (representing a more trustworthy result of the intervention) (Gorard, 2014). The sieve contains six categories: design, scale, dropout, outcomes, fidelity, and validity, with the overall star rating (or evaluation) corresponding to the lowest classification received in any of the categories (Gorard, 2014). If the rating for a particular category cannot be determined from the reports, that category will automatically be assigned a rating of zero (Gorard, 2014). The same second reviewer, mentioned in the selection process, will screen a random 10% sample of the quality assessments. Since no established guidelines are available for each rating cutoff in relation to Scale, the author and her dissertation supervisor agreed on the following criteria to ensure consistent ratings for the purpose of this review:

Scale: Number of participants per comparison group

- Very small: $n \leq 10$
- Small: $10 < n \leq 20$
- Medium: $20 < n \leq 50$
- Large: $n \geq 50$

Dropout:

- Minimal attrition: 0-10% dropout (over 90% completion)
- Some attrition: 11-20% dropout (between 80-90% completion)
- Moderate attrition: 21-30% dropout (between 70-80% completion)
- High attrition: over 30% dropout (less than 70% completion)

Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics*, *110*, 47-60. http://www.radstats.org.uk/no110/Gorard110.pdf

Describe how risk of bias, trustworthiness, or quality of individual studies will be assessed. Name any specific tools, e.g. Gorard's Sieve, Maryland Scientific Methods Scale, Cochrane Risk of Bias Tool, EPPI Weight of Evidence Tool, etc. State how this information will be used in the synthesis.

Data Synthesis

A meta-analysis will be performed if there is sufficient data with comparable outcomes in the literature. If the research is limited or varies widely in outcome measures, a narrative synthesis will be used instead. A narrative synthesis provides "a summary of the current state of knowledge" to address the research question(s) (Popay et al., 2006, p. 6). This approach enables the integration of findings from the included studies, analysis of patterns within the data, and evaluation of the strength of evidence (Popay et al., 2006).

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K. & Duffy, S., 2006. Guidance on the conduct of narrative synthesis in systematic reviews. *A product from the ESRC methods programme Version*, *1*(1), p.b92.

Describe criteria under which quantitative synthesis will be performed. If quantitative synthesis is appropriate, describe preferred summary measure (Cohen's D, Hedges' G, etc.) and how these will be combined. Describe any additional planned analyses (e.g. sub-group analysis). If quantitative synthesis is not appropriate, describe how data will be combined and summarised.

Meta-biases

A funnel plot will be created to assess publication bias, assuming the eligible studies report sufficient data and are comparable to warrant a meta-analysis.

Describe how meta-biases (publication bias, selective outcome reporting, etc.) will be addressed.

Confidence in cumulative evidence

The GRADE framework will be used to assess the level of confidence in cumulative evidence, taking into account five key factors: (1) risk of bias; (2) inconsistency in findings; (3) indirectness of evidence; (4) imprecision (indicated by the width of confidence intervals); and (5) publication bias.

Describe how the strength of the body of evidence will be assessed.

Sources of Funding

N/A

Specify any financial or other support for the review. Provide name for the review funder and/or sponsor. Include funding reference number if available.

Role of Funders

N/A

Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol.

Anticipated or actual start date: January 25, 2025

Anticipated completion date: September 30, 2025

Other language resources

N/A

If you have other language versions of your protocol, please include a link here.

Current Status

☒ Ongoing
☐ Completed but not published
☐ Completed and published

For initial submissions select 'ongoing'. When you have completed the review, you should return to update the record accordingly. Your review status will be displayed next to your record in IDESR.

Details of Published Review

N/A

Once the review is complete, you should return to this form to add information about where it has been published. Include the full bibliographic reference and a link to the published document, a DOI or URL. You should link to a pre-print as soon as one is available, then update this information once the full review has been published. Your review will then be added to the IDESR database.

Appendix B: Boolean strings for individual databases

Database 1: Web of Science
(“Chinese as a foreign language” OR “Mandarin as a foreign language” OR “Chinese character*” OR “Chinese orthograph*” OR “Chinese writing system” OR “Chinese learning”)

AND (“Chinese reading” OR “Chinese writing” OR “Chinese literacy” OR “radical knowledge” OR “character recognition” OR “character acquisition” OR “character knowledge” OR “character retention” OR “character recall” OR “character improvement” OR reading development)

The search was additionally restricted to the following categories:

- Education Educational Research
- Linguistics
- Language Linguistics
- Social Sciences Interdisciplinary

Database 2: Linguistics Collection (including LLBA)
noft("Chinese reading" OR "Chinese writing" OR "Chinese literacy" OR "radical knowledge" OR "character recognition" OR "character acquisition" OR "character knowledge" OR "character retention" OR "character recall" OR "character improvement")

AND noft(“Chinese as a foreign language” OR “Mandarin as a foreign language” OR “Chinese character*” OR “Chinese orthograph*” OR “Chinese writing system” OR “Chinese learning”)

Database 3: ProQuest Social Science Premium Collection (including ERIC)
noft("Chinese reading" OR "Chinese writing" OR "Chinese literacy" OR "radical knowledge" OR "character recognition" OR "character acquisition" OR "character knowledge" OR "character retention" OR "character recall" OR "character improvement")

AND noft(“Chinese as a foreign language” OR “Mandarin as a foreign language” OR “Chinese character*” OR “Chinese orthograph*” OR “Chinese writing system” OR “Chinese learning”)

Database 4: ProQuest Dissertations & Theses Global
noft("Chinese reading" OR "Chinese writing" OR "Chinese literacy" OR "radical knowledge" OR "character recognition" OR "character acquisition" OR "character knowledge" OR "character retention" OR "character recall" OR "character improvement")

AND noft(“Chinese as a foreign language” OR “Mandarin as a foreign language” OR “Chinese character*” OR “Chinese orthograph*” OR “Chinese writing system” OR “Chinese learning”)

Database 5: SCOPUS
noft("Chinese reading" OR "Chinese writing" OR "Chinese literacy" OR "radical knowledge" OR "character recognition" OR "character acquisition" OR "character knowledge" OR "character retention" OR "character recall" OR "character improvement")

AND noft(“Chinese as a foreign language” OR “Mandarin as a foreign language” OR “Chinese character*” OR “Chinese orthograph*” OR “Chinese writing system” OR “Chinese learning”)

Database 6: PsycINFO
noft("Chinese reading" OR "Chinese writing" OR "Chinese literacy" OR "radical knowledge" OR "character recognition" OR "character acquisition" OR "character knowledge" OR "character retention" OR "character recall" OR "character improvement")

AND noft(“Chinese as a foreign language” OR “Mandarin as a foreign language” OR “Chinese character*” OR “Chinese orthograph*” OR “Chinese writing system” OR “Chinese learning”)

Database 7: British Education Index
noft("Chinese reading" OR "Chinese writing" OR "Chinese literacy" OR "radical knowledge" OR "character recognition" OR "character acquisition" OR "character knowledge" OR "character retention" OR "character recall" OR "character improvement")

AND noft(“Chinese as a foreign language” OR “Mandarin as a foreign language” OR “Chinese character*” OR “Chinese orthograph*” OR “Chinese writing system” OR “Chinese learning”)

Appendix C: Data extraction form for Study 2

Administrative Information

Full citation: Chen, R. (2020). *The use of UMU interactive platform in Chinese language reading skills of Grade 10 Thai students*. [Master's thesis, Rangsit University]. Thailand.

What date was the form completed?	Saturday 1 March 2025
Publication type *(e.g. thesis, journal article, book chapter)*	Master's thesis
Language	English
Database source *(e.g. SCOPUS, PsycINFO)*	ProQuest Dissertations & Theses Global
Source of funding	Not stated

Contextual Information

When was the study conducted?	January-February 2020
Study/research location	Rayong, Thailand
Research question(s)	Would UMU interactive platform improve Chinese language reading skills of Grade 10 Thai students?
Educational level *(e.g. primary, secondary, tertiary)*	Secondary (Grade 10)
Type of learning *(e.g. language program, part of school or tertiary curriculum)*	Liberal Arts' students studying Chinese as a foreign language
Type of institution *(e.g. public school, elite university)*	Private school
Delivery *(e.g. online, in-person)*	In-person and online components
Writing script *(e.g. traditional or simplified)*	Simplified
Participant demographics *(e.g. age, gender, L1 background, L2 proficiency, socio-economic background)*	Age: 16-17 years Gender: 6 females and 6 males L1 background: Thai CFL proficiency: Participants had learned Chinese for 3-6 years. Most had master essential listening, speaking, reading, and writing skills. All Chinese lessons in school were taught by native Mandarin speakers and Thai teachers. The class was mixed-ability.

Intervention Information

Study/research design *(e.g. quasi-experimental, randomised controlled trial)*	Single-group design
Study duration	One month
Number and sample size of comparison groups *(including dropout, descriptions of treatment and control groups)*	Grade 10 students (n=12). No comparison group. No participants dropped out (100% completion rate).
Method of participant allocation *(e.g. random, in-tact classes, prior scores, unknown/not reported)*	Intact classes
Measurement tools *(e.g. standardised tests used if mentioned, descriptions and numbers of tests, were they piloted or validated and/or blind-marked?)*	Tests were designed by the author and followed the structure of the standardised reading test: Hanyu Shuiping Kaoshi (Level 2). Tests were not piloted but validated by an Associate Professor at Rangsit University and two native Chinese teachers (one from Burapha University; and the other from Assumption College Rayong).
Teaching approach(es) *(descriptions of the lesson(s), including the instructors involved and their experiences or background, duration of each lesson, was it compared to other pedagogical approaches?)*	The author was the instructor. He developed four lesson plans (100-minutes each). There were eight lessons in total (one lesson plan = two teaching sessions). Pupils attended class twice per week. 1. Students watched online videos on the learning platform and studied key words/grammar before attending in-person classes. They practised reading skills online, under the guidance of the teacher. 2. Students completed homework using the platform. Lesson plans: The figure originally presented here cannot be made freely available via ORA because of copyright. The figure was sourced at *The use of UMU interactive platform in Chinese language reading skills of Grade 10 Thai students.*

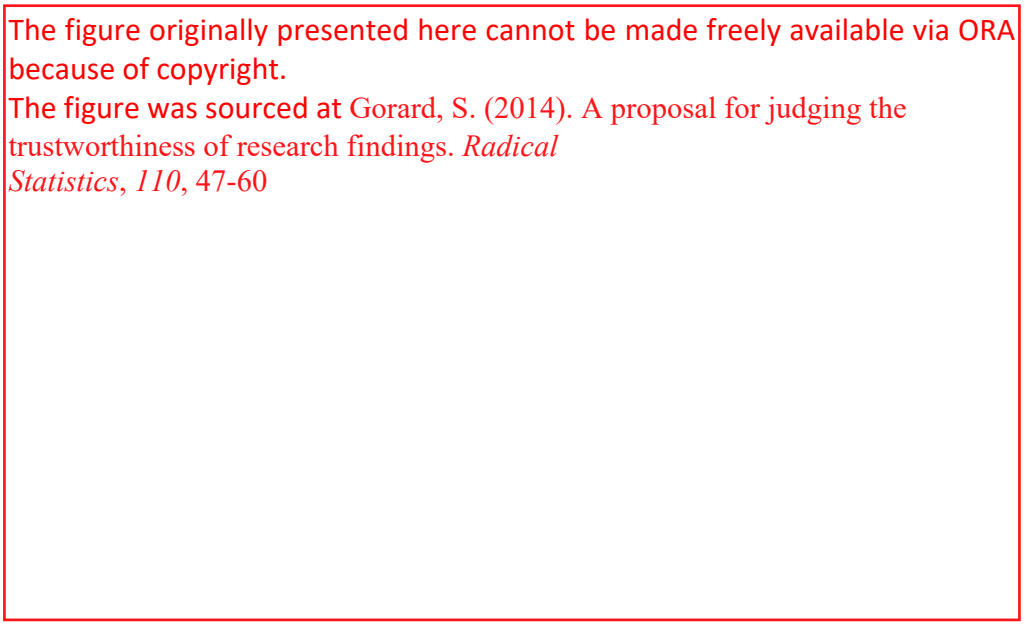
	The figure originally presented here cannot be made freely available via ORA because of copyright. The figure was sourced at *The use of UMU interactive platform in Chinese language reading skills of Grade 10 Thai students.*
Teaching and learning materials *(e.g. vocabulary lists, flashcards, character worksheets with stroke sequences)*	No examples of teaching/learning materials were provided.

Outcome Information

What language skills or types of reading outcomes were assessed? *(e.g. reading comprehension and/or*	Reading comprehension

fluency, character knowledge and retention, etc.)	
Pre-post- test measure(s)	Pre-post-tests contained identical questions/content. Pre-tests were conducted online via the learning platform, while post-tests were completed via pen-and-paper. Tests were scored out of 25 (one mark per question). Assessment tests included four sections: 1. Sentence-picture matching 2. Fill-in-the-blanks with options provided 3. Students read passages/conversations and answer true/false questions based on what they read 4. Match questions with its corresponding answers
How was data analysed? *(e.g. names of statistical software)*	Data were analysed using the Wilcoxon Signed Rank Test.
Effect size according to Cohen's D or Hedges' G	Not stated
Findings *(e.g. mean scores, standard deviations, outcomes of pre-post-intervention)*	The figure originally presented here cannot be made freely available via ORA because of copyright. The figure was sourced at *The use of UMU interactive platform in Chinese language reading skills of Grade 10 Thai students.*
Author(s) conclusions	The interactive learning platform has a significantly positive effect on reading performance. It helps to improve retention and promotes self-study.

Appendix D: Copy of Gorard's (2014) sieve



Due to the absence of established guidelines for rating cutoffs in Scale and Dropout, the reviewer in collaboration with their supervisor, developed the following criteria to maintain consistency in the ratings for the purposes of this review:

Scale:

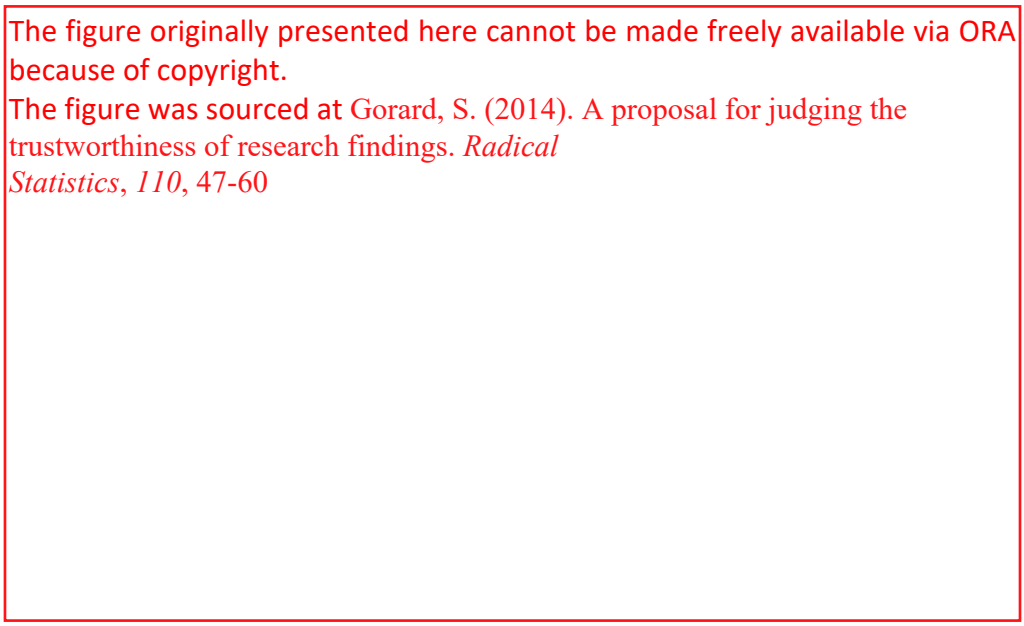
- Trivial: $n \leq 5$
- Very small: $5 < n \leq 15$
- Small: $15 < n \leq 20$
- Medium: $20 < n \leq 50$
- Large: $n \geq 50$

Dropout:

- Minimal: 0-10% dropout (over 90% completion)
- Some: 11-20% dropout (between 80-90% completion)
- Moderate: 21-30% dropout (between 70-80% completion)
- High: Over 30% dropout (less than 70% completion)

Appendix E: RoB assessment for Study 2

Full citation: Chen, R. (2020). *The use of UMU interactive platform in Chinese language reading skills of Grade 10 Thai students*. [Master's thesis, Rangsit University]. Thailand.



Appendix F: References of included studies

S1: Chang, L. Y., Xu, Y., Perfetti, C. A., Zhang, J., & Chen, H. C. (2014). Supporting orthographic learning at the beginning stage of learning to read Chinese as a second language. *International Journal of Disability, Development and Education*, *61*(3), 288-305.

S2: Chen, R. (2020). *The use of UMU interactive platform in Chinese language reading skills of Grade 10 Thai students*. [Master's thesis, Rangsit University]. Thailand.

S3: Chiu, C. (2024). *How does writing instruction impact Chinese as a foreign language learners' literacy?*. [Doctoral thesis, University of Delaware]. United States of America.

S4: Chung, K. H. (2007). Presentation factors in the learning of Chinese characters: The order and position of Hanyu Pinyin and English translations. *Educational Psychology*, *27*(1), 1-20.

S5: Harvey, R. E., & Brooks, P. J. (2022). Effects of text messaging using digital Pinyin input on literacy skills of elementary Chinese immersion learners. *Language Teaching Research*, *0*(0), 1-27.

S6: He, J., & Huang, H. (2014). Learning Chinese characters with animated etymology. *International Journal of Computer-Assisted Language Learning and Teaching*, *38*(1), 99-118.

S7: Ke, S., & Dubravac, S. (2021). When should characters be introduced to novice-level Chinese in a blended learning setting?. *Studies in Chinese Learning and Teaching*, *6*(1), 1-26.

S8: Knell, E., & West, H. (2017). To delay or not to delay: The timing of Chinese character instruction for secondary learners. *Foreign Language Annals*, *50*(3), 519-532.

S9: Li, W. (2004). *The "grapheme combination method": Teaching and learning Chinese characters through associative links*. [Doctoral thesis, George Mason University]. United States of America.

S10: Li, J. T., & Tong, F. (2020). The effect of cognitive vocabulary learning approaches on Chinese learners' compound word attainment, retention, and learning motivation. *Language Teaching Research*, *24*(6), 834-854.

S11: Osborne, C. (2016). Chinese in the classroom: Initial findings of the effects of four teaching methods on beginner learners. *Journal of Second Language Teaching and Research*, *5*(1), 202-225.

S12: Osborne, C. (2018). Examining character recognition and recall skills of CFL beginner learners under four different approaches. *Journal of the Irish Association for Applied Linguistics*, *25*(1), 52-73.

S13: Osborne, C., Zhang, Q., & Zhang, G. X. (2020). Which is more effective in introducing Chinese characters? An investigative study of four methods used to teach CFL beginners. *Language Learning Journal*, *48*(4), 385-401.

S14: Osborne, C., Zhang, Q., & Adamson, B. (2022). The next steps for teaching characters in CFL: Investigating the effects of four Character teaching methods on beginner learners. *International Journal of Chinese Language Education*, *11*(1), 45-82.

S15: Packard, J. L. (1990). Effects of time lag in the introduction of characters into the Chinese language curriculum. *Modern Language Journal*, *74*(1), 167-175.

S16: Poole, F., Clarke-Midura, J., & Ji, S. (2022). Exploring the affordances and effectiveness of a digital game in the Chinese dual language immersion classroom. *Journal of Technology and Chinese Language Teaching*, *13*(1), 46-73.

S17: Poole, F., & Sung, K. (2015). Three approaches to beginning Chinese instruction and their effects on oral development and character recognition. *Eurasian Journal of Applied Linguistics*, *1*(1), 59-75.

S18: Ren, G. (2004). Introducing OVAL writing: A new approach to Chinese character retention for secondary non-Chinese-speaking background learners. *In Australian Federation of Modern Language Teachers Associations National Conference*. https://search.informit.org/doi/10.3316/aeipt.138737

S19: Shen, H. H. (2010). Imagery and verbal coding approaches in Chinese vocabulary instruction. *Language Teaching Research*, *14*(4), 485-499.

S20: Taft, M., & Chung, K. (1999). Using radicals in teaching Chinese characters to second language learners. *Psychologia*, *42*(4), 243-251.

S21: Tsai, C. H. (2014). *Multimedia mediation and Chinese orthographic character learning among non-heritage CFL beginners*. [Doctoral thesis, University of Iowa]. United States of America.

S22: Wang, L. (2005). *The impact of multimedia on Chinese learners' recognition of characters: A quantitative and qualitative study*. [Doctoral thesis, Purdue University]. United states of America.

S23: Wang, T. (2024). Designing a digital game for Chinese character learning: A theory-driven practice approach. *Education Sciences*, *14*(12), 1-14.

S24: Wu, Y. (2012). Using external text vocalisation to enhance reading development among beginning level Chinese learners. *Journal of the Chinese Language Teachers Association*, *47*(1), 1-23.

S25: Xu, P., & Jen, T. (2005). “Penless” Chinese language learning: A computer-assisted approach. *Journal of the Chinese Language Teachers Association*, *40*(2), 25-42.

S26: Xu, X., & Ke, F. (2020). Embodied interaction: Learning Chinese characters through body movements. *Language Learning & Technology*, *24*(3), 136-159.

S27: Xu, X., & Padilla, A. M. (2013). Using meaningful interpretation and chunking to enhance memory: The case of Chinese character learning. *Foreign Language Annals*, *46*(3), 402-422.

S28: Xu, Y., Chang, L. Y., Zhang, J., & Perfetti, C. A. (2013). Reading, writing, and animation in character learning in Chinese as a Foreign Language. *Foreign Language Annals*, *46*(3), 423-444.

S29: Xu, Y., Chang, L. Y., & Perfetti, C. A. (2014). The effect of radical-based grouping in character learning in Chinese as a foreign language. *Modern Language Journal*, *98*(3), 773-793.

S30: Zhang, T. (2019). *Online comics for the teaching and learning of Chinese language in the Australian context*. [Master’s thesis, Western Sydney University]. Australia.