Xuan Yang and Caitríona Osborne*

The development and validation of a C-test and a pseudo-character test for online CFL assessment

https://doi.org/10.1515/jccall-2022-0019
Received August 19, 2022; accepted May 8, 2023; published online June 30, 2023

Abstract: Taking inspiration from a popular online English test, this study first documents the development of an achievement test containing a newly developed C-test and pseudo-character test for beginner-level Chinese learners. Then, based on the test results collected from 53 university learners and using statistical tools, analyses were conducted on its content validity, construct validity, criterion-related validity, difficulty, discrimination power, and reliability. Results indicate that this new achievement test has relatively satisfactory reliability and validity, with only minor issues that need to be addressed in future studies. Situated in the context of language assessment, the study sheds light on the application of the C-test and a pseudo-character section in both face-to-face and at-distance Chinese as a foreign language (CFL) classrooms. Furthermore, the study also provides practical and applicable methods for CFL teachers to develop their own assessments.

Keywords: C-test; Chinese assessment; Chinese characters; pseudo-word test; teaching Chinese as a foreign language

1 Introduction

After the Covid-19 pandemic began in early 2020, universities worldwide rapidly transferred all learning online in an attempt to stop the spread of the virus. Naturally, this switch to remote learning had a serious effect not only on the learning environment in universities but also on in-person language proficiency test centres. While online tests allow test takers around the world to access such at their own convenience, the tests are not without criticism. Rather than re-designing the test paper for a more efficient experience, the format and content of the online Hanyu

***Corresponding author: Caitríona Osborne**, University College Dublin, Dublin, Ireland, E-mail: caitriona.osborne@ucd.ie
Xuan Yang, University College Dublin (UCD) Confucius Institute for Ireland, Dublin, Ireland / Renmin University of China, Beijing, China, E-mail: xuan.yang@ucd.ie, cucyangxuan@gmail.com. https://orcid.org/0000-0002-1189-7066

Shuiping Kaoshi (HSK) 2.0 followed the same design as its paper version, only with additional invigilation measures such as setting up of a camera behind each test taker (Chinese Testing International, 2022a). Test takers must wait for a test date and their exam results take a minimum of 10 days to arrive (ibid.). In addition, there is a disparity between online and on-paper results, with people taking the test online generally scoring higher. Two reasons given in Peng et al.'s paper (2021) for this disparity in scores are that (1) online test takers can use headphones and therefore hear the listening tasks more clearly and (2) online test takers' writing method is Pinyin input rather than handwriting. Given these issues and the slow pace of change, it is important to experiment with new types of questions for a Chinese test in the digital age.

With the likes of the Duolingo English Test (DET)'s popularity and endorsement by over 4,000 institutions worldwide (Duolingo, 2023), coupled with a efficiency in cutting testing time by over a third compared to other standardised English proficiency tests (Wagner, 2020), there seems to be more acceptance of such worldwide. While not perfect, it is worth exploring whether the design of the DET could inform a new Chinese test for improved efficiency and user experience in the digital age, particularly in reference to the C-test – a variation of the cloze test – and the pseudo-word test – a type of Yes-No Vocabulary Task.

Despite technological advances, producing, developing, and analysing tests remain fundamental. The present study therefore scrutinises the newly developed C-test and a pseudo-character section for measuring CFL beginners' linguistic proficiency. It particularly focuses on the validity of these two question types before they are used online. While it is common for teachers to develop tests to measure student learning outcomes (Zhang, 2017), due to time constraints and inappropriate content and/or methods, these teacher-created tests often experience measurement errors resulting in reduced validity or reliability (Barrette, 2004). Thus, it is also essential to provide CFL teachers with practical guidelines for developing offline and online tests, which is another feature of this paper.

2 Literature review

2.1 Online foreign language assessment

Although remote teaching and testing have come under the spotlight in recent years as a result of the emergency response to the Covid-19 pandemic, online or computer-assisted testing in the foreign language classroom is not necessarily a new phenomenon. However, the mass migration of courses and tests to the online space in

recent years, accelerated by Covid-19, warrants further research as the trend in at-distance and/or blended courses continues to rise.

Indeed, online tests can come in the form of international proficiency tests, university exams, or even in-class quizzes and assessments. With advances in technology in general, language teachers and instructors worldwide incorporate online assessment tools (such as *Quizlet*, *Kahoot*) to assist student learning, which are reported to be motivating and beneficial to both teachers and students in the foreign language classroom (e.g., Huyen, 2022). Certainly, such online tools are particularly beneficial when assessing a language that has limited opportunities to practise in a given country (Alsied & Pathan, 2013), while a study conducted by VanPatten et al. (2015) demonstrated no difference in scores between an in-class and online test in a particular communicative and proficiency-oriented language course. Overall, using online tools to assess foreign languages has proven to be convenient, reliable, and efficient (e.g., Wagner, 2020).

However, during the development of emergency online CFL assessments during the Covid-19 pandemic, serious issues relating to online assessment design emerged such as: an increase in opportunity for students to cheat (Daniels et al., 2021); the need to generate anti-OCR (Optical Character Recognition) documents for assessment (Wang & East, 2020); issues in testing handwriting (ibid.); and both students' and teachers' technology skills (Burns et al., 2020; Jin et al., 2021).

Indeed, one of the main issues in online CFL testing lies in the difficulty of testing handwriting and therefore character formation. Certainly, typing was seen to be used in online testing during the pandemic (e.g., Wang & East, 2020), however, as Guan et al. (2011) demonstrate, typing is not an appropriate substitute for assessing character formation. In addition, there is a paucity of research in the area of online CFL assessment methods, despite a huge increase of online CFL teaching and learning triggered by Covid-19 (e.g., Wang & East, 2020; Zhang, 2020). As a result, it is vital to go back to the basics of language assessment development to ensure that a successful online CFL assessment can be developed and validated.

2.2 Language assessment development

According to Zhang (2016), the process of developing a language test has three stages: (1) design and planning; (2) execution and administration; and (3) post-test analysis.

Understanding examinees is important for developing language assessments (Zhang, 2016), while Bachman (1990) prioritizes the test's purpose in test design. To achieve the predetermined objectives, it is vital to select an appropriate exam type (e.g., proficiency tests, diagnostic tests, achievement tests) (Brown, 2004). The number of items and scoring procedure can also be significant factors since they

influence the weighting of content (ibid.). Item writing and delivering the test are the main steps of test execution and administration. Alderson et al. (1995) note a good test writer should have specific experience in teaching and assessing students in the area being tested, while creativity is also essential as poorly constructed questions can damage the alignment of the assessment process (Tavakol & Dennick, 2011). Post-exam analysis techniques are then used to improve the quality and reliability of assessments (ibid.). Statistical analysis, item analysis, validity, and reliability are all important factors in evaluating a newly designed test, with the validity test being the most significant.

Although validity is generally viewed as a unified concept (Bachman, 1990; Messick, 1989), it is more practical to see validity as segmented into subcategories or steps that may easily be approached with manageable tools (Dobrić, 2018). The major and complementary types of evidence that need be gathered during the validation process include construct validity, content validity, and criterion-related validity (Bachman, 1990; Weir, 2005). In achieving criterion validity, an older, well-established test is typically used as a criterion, as it also serves as an indicator of the ability being tested (Bachman, 1990; Weir, 2005). The result of the comparison is usually expressed as a correlation coefficient, with higher coefficients demonstrating closely related and reliable tests (Alderson et al., 1995). Content validity is defined as "a conceptual or noun statistical validity based on a systematic analysis of the test content to determine whether it includes an adequate sample of the target domain to be measured" (Davies et al., 1999, p. 34). To establish content validity, experts are called upon to examine or rate the test tasks or items with a data collection instrument (Alderson et al., 1995; Feng et al., 2020). Conducting construct validation involves testing hypothesized relationships between test scores and abilities empirically (Bachman, 1990). Alderson et al. (1995) introduces five different approaches to achieve construct validity, including confirmatory factor analysis (CFA).

2.3 Language assessment development in the context of CFL

In the context of CFL, language assessment development requires deep consideration given the unique features of the language, particularly in the written form. The Chinese writing system is logographic and morpho-syllabic, whereby characters correspond to both syllables and morphemes (DeFrancis, 1989). This makes it difficult for L2 learners to determine whether new characters are formed based on pronunciation, meaning, or both (Osborne et al., 2018). Characters are also composed of radicals which can be further decomposed into strokes (Liu et al., 2007). Previous studies have shown that stroke patterns (Chen et al., 1996), and radicals (Anderson et al., 2013) both can function as units of character perception. For CFL students, it is

imperative to be able to recognise Chinese characters as a whole as well as the identifiable units within, meaning that assessment design is vital in the CFL classroom.

The HSK 2.0 test, introduced in 2010, is China's standardised language proficiency examination and is administered globally (McNaughton, 2005). According to Chinese Testing International (2022b), its main purpose is to assess Chinese language ability in various contexts, including daily, academic, and professional life. The HSK 3.0 has been introduced in 2021 with reforms in vocabulary distribution (three levels with nine bands) and the introduction of translation and handwriting on the test paper (China Education Center, 2023). While the participants of the current research are registered to a module aligned to HSK 2.0, with the introduction of handwriting exercises in the HSK 3.0 test paper, the need for the current research is again highlighted, particularly in reference to the online space.

In addition to nationally administered tests, several studies have focused on four language skills – reading, writing, speaking, and listening – and two types of language knowledge – vocabulary and grammar (Zhang, 2017). Given the unique features of Chinese, Chinese language teachers and researchers should pay more attention to developing new types of test questions that specifically target the unique features of Chinese. However, as Zhang (2017) notes, the majority of studies in developing Chinese testing are published in Chinese, making it difficult for the international community to keep abreast of developments in the field. Furthermore, Zhang (2017) also concludes that there is a paucity of literature on the assessment of reading Chinese – including the involvement of Pinyin on test papers – as well as in assessing character writing with computers. To address these limitations, we have identified that two test types – C-test and pseudo-character – are particularly promising.

While the cloze test has been demonstrated to be a valid instrument for measuring Chinese language proficiency (Feng et al., 2020) and the C-test, a variation of the cloze test, is predominantly used in English tests, the C-test is yet to be reported on in the context of teaching/learning CFL. Second, due to the fact that CFL students are generally not equipped with adequate amounts of characters (Zhou, 2007), as well as experiencing considerable difficulties identifying, writing, and using characters (Wu et al., 2017), educators should prioritise the development of tests that can assess character knowledge. The pseudo-word test may therefore have the potential to be modified as a pseudo-character test for the CFL context.

2.4 C-test

The cloze test was first introduced in the 1950s (Taylor, 1953) as a method of language testing (Oller, 1972, 1973). A considerable number of studies have demonstrated that

the cloze test can not only serve as a standardised test to measure students' language proficiency (Alderson, 1979; Feng et al., 2020; Oller, 1973) but also has a wide range of correlation coefficients with well-established tests such as TOEFL and HSK (Bachman, 1985; Feng et al., 2020). A number of variations of the cloze test have been developed over the past few decades (Zhang et al., 2005), one being the C-test developed by Raatz and Klein-Braley in the 1980s (Raatz & Klein-Braley, 1981).

While making the overall test shorter, the developers of the C-test "damage" words rather than delete them (Klein-Braley, 1997, p. 64). As an example of a C-test adapted from the DET, Figure 1 shows a paragraph that is presented to the test takers, with certain words missing letters. The test taker must be able to understand the context and spell the words correctly in order to pass this section.

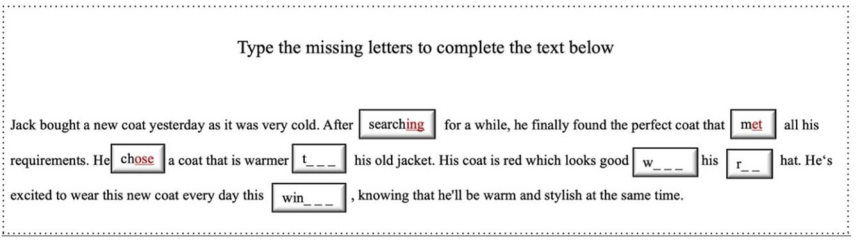


Figure 1: Sample of C-test adapted from Cardwell et al. (2022).

As a stable, economical, reliably scoring, and easy-to-write/administer test, the C-test has been widely used by test developers (Grotjahn, 2002; Harsch & Hartig, 2016). Compared with other reduced redundancy tests, the C-test comes top in terms of difficulty level, reliability, validity, and factorial validity (Klein-Braley, 1997). Additionally, given the nature of the C-test design, the answers are more definite, rather than being subjective as in cloze tests (ibid.). Similarly, the frequency of damaged words allows more classes of words to be tested (ibid.). Besides being highly correlated with listening and speaking tests, the C-test has also been shown to be used to measure students' overall language ability (Eckes & Grotjahn, 2006). Therefore, as long as reasonable adaptations are made, the C-test may also be a useful tool for assessing Chinese learners holistically across phonetics, vocabulary, grammar, and discourse.

2.5 From pseudo-word test to pseudo-character test

The Yes-No Vocabulary Test (YNVT) format is widely used by language learners due to its efficiency, easy item development, and straightforward scoring system (Nation,

1990). However, one of the drawbacks of YNVT is students' guessing behaviours (Pellicer-Sánchez & Schmitt, 2012). To address this issue, the pseudo-word test, developed by Zimmerman et al. (1977), is commonly used. In such sections, test takers are presented with a set of written real words mixed with pseudo words designed to appear like real words, and must choose the correct ones (Cardwell et al., 2022) (see Figure 2). In the context of English, the pseudo-word test can predict listening, reading, and writing skills (Milton et al., 2010).

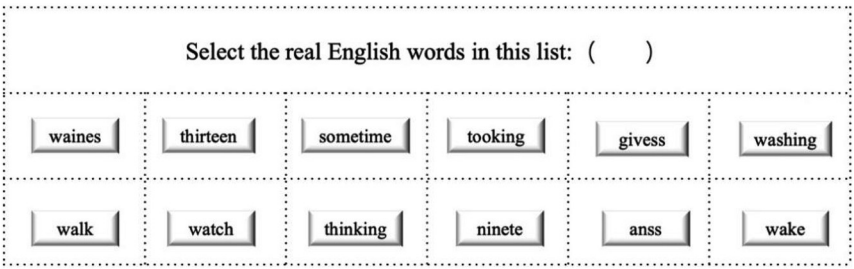


Figure 2: Sample of written YNVT pseudo-word test adapted from Cardwell et al. (2022).

In the context of CFL, pseudo-word tests have been adapted to pseudo-character tests. It is widely agreed that pseudo-characters serve as valuable stimuli in the exploration of the cognitive processes underlying character recognition of Chinese readers. For instance, the pseudo-characters can be used as testing items for monitoring the progress of Chinese learners' orthographic awareness (Chung et al., 2012). Chang et al. (2022) developed a Chinese pseudo-character/non-character producing system (CPN) based on traditional Chinese characters and the results of their study demonstrate that highly proficient Chinese readers are better able to identify pseudo-characters. However, using pseudo-characters has also been subject to some criticisms. First, there is no consensus among researchers regarding the definition of pseudo-characters. Chen et al. (1996) defined pseudo-characters as the swapping of two real characters and the creation of two non-existent combinations of orthographic units that always display lexical radicals in their legitimate positions. Similarly, Sze et al. (2014) created pseudo-characters by swapping semantic components within each pair while Kuo et al. (2015) created pseudo-characters that resemble real Chinese characters in structure but do not exist in Chinese by combining a high-frequency semantic radical with one or more simple characters. In addition, Shi (2000) categorized students' Chinese character writing errors as non-characters (非字), fake characters (假字), and characters with errors (别字), which can resemble pseudo-characters that are artificially created. However, numerous

studies have shown that presenting errors to students in the long term can lead to anxiety, confusion, and reinforcement of errors (Khansir & Pakdel, 2018). Lastly, with a lack of tools for precisely creating pseudo-characters, it is a time-consuming task to create such (Chang et al., 2022).

While there has been some research into pseudo-characters in the context of CFL, very little academic research has been conducted on whether pseudo-characters could be used for achievement tests, how well they work, and the potential issues. This study thus aims to develop a simplified Chinese pseudo-character test for a Chinese beginner's achievement test and address any issues therein.

The literature review has revealed some gaps in developing CFL assessment. First, the paucity of research in the area of Chinese language assessment development, particularly published in English, demonstrates the need for the present study. CFL teachers need valid and reliable testing instruments as well as essential knowledge on assessment development. In addition, C-test and pseudo-character tests designed for CFL students and teachers are lacking. Finally, innovative assessment methods that can also be applied to the online space should be explored in the CFL context. This study therefore aims to fill these gaps by addressing the following research questions:

RQ1: How can a C-test and pseudo-character test be developed for assessment in the context of CFL?

RQ2: How accurately can the newly developed C-test and pseudo-character test assess students' learning outcomes?

3 Methodology

3.1 Test type and scope

The current study develops an achievement test that contains two sections: (1) a C-test section and (2) a pseudo-character section for a 12-week (36 teaching hours) module. This module covers content from the absolute beginner level and caters to students with very little or no prior knowledge of Mandarin Chinese. The HSK Standard Course Textbook Volume 1 (Jiang, 2014) is the main material used. Upon successful completion, participants have acquired basic communicative competence and knowledge of pronunciation and tones, basic sentence structures, useful daily

expressions, core vocabulary, and high-frequency-appearing characters broadly aligned to the HSK 1 test.

3.2 Item writing for C-test section

The C-test comprises 282 Chinese characters with 15 blanks (see Figure 3). All characters were selected from the HSK 1 textbook and supplementary materials used by the teachers during the module. Words that are not required to be recognised by participants, such as 西班牙 (Spain), are prompted with Pinyin.

A. Complete the text below by filling in the blanks. (1*15=15)

•③⑦⑫ Please complete the characters with the correct radicals/components.

•①⑭ Please fill in the Pinyin according to the characters provided.

•②④⑤⑥⑧⑩⑪⑮ Please fill in the blank with ONE character ONLY.

•⑨⑬ Please fill in the blank with NO MORE THAN FIVE characters.

大家好!我叫李月. 我今年二十三(岁①_____)了, 我的生日是一(②_____)二十三号。我是法国人。我(③__豕__) 有五口人:妈妈,爸爸,一个妹妹、一个弟弟和我。我的妈妈是美国人,爸爸是爱尔兰人。我星期二晚上在大学(④_____)习汉语。我会读汉语,(⑤_____) 会写汉字,我非常喜欢汉字。我(⑥_____)医院工作,我是一个医生。

现在是早上八点半,我在看书。我妈妈在给奶奶(⑦扌 _____)电话,妹妹(⑧_____) 睡觉,爸爸在吃早饭。我的爸爸很(⑨_____)吃水果。今天早上 ,他吃了很多苹果。我的弟弟今天(⑩_____) 在家,他昨天和他的朋友一(⑪_____) 去了西班牙(Xībānyá), 他们是坐飞(⑫木_____) 去的。西班牙的(⑬_____)很好,不冷也不(热⑭ _____)。弟弟和他的朋友都很爱吃西班牙菜。他们去了很多西班牙饭店,也认识(⑮_____)很多西班牙朋友。

Figure 3: The C-test developed for the current study.

The researchers not only attempted to provide definite answers when writing C-test items, but also more detailed and eye-catching instructions, such as "②④⑤⑥⑧⑩⑪⑮: Please fill in the blank with ONE character ONLY".

In addition, the current study adapted the concept of "damaged" items (家, 打, and 机). Specifically, one or two components of the chosen characters are deleted, and participants are asked to fill in the missing components. For example, participants are asked to complete the character 家 (family) in the following sentence: "我(③__豕__)有五口人 (there are five people in my family)." Furthermore, a similar approach is used in the testing of Chinese words whereby researchers "damaged" the

words by deleting one character and asked the participants to fill in the missing characters. For example, the participants are required to provide the missing character "起" in the following sentence: "他昨天和他的朋友一(⑪___) 去了西班牙(Xībānyá) (He went to Spain with his friends yesterday)."

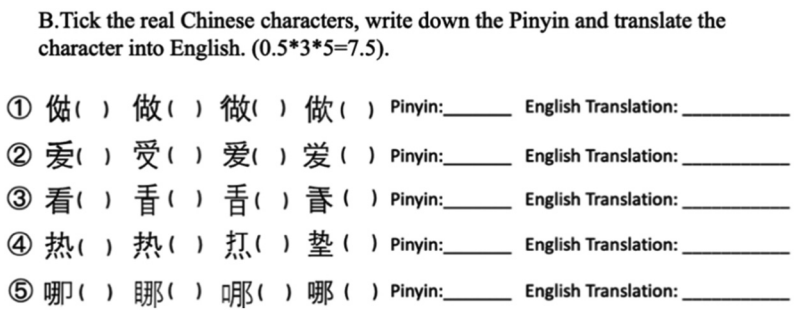
Most CFL teachers teach phonology when students first begin learning Chinese, but do not continue doing so afterward (Ye, 2003). Thus, incorporating a pinyin test into an achievement test for CFL beginners is essential. The initials "s" and "r", as well as the finals "ui" and "e", which have been shown by studies to be easily mistaken by students (Ran & Yu, 2019) are examined. Students are required to write the pinyin according to the characters " 岁" and " 热" indicated in brackets in question ① and question ⑭.

In fact, the content areas in the newly designed C-test correspond to a range of learner proficiencies from basic (e.g., phonology, character, morphology, and syntax) to more advanced (e.g., cohesion and grammar). The morpheme and word categories contained four items testing verbs, prepositions, adverbs, and concrete nouns, representing 26.66 % of the C-test. Regarding grammar, progressive action, complex sentences, and the particle "了" were included. Additionally, some of the questions required students to fill out the correct answers by understanding the contextual discourse logically. For example, in the sentence "我的爸爸很(⑨____)吃水果", only students who understand the meaning of the sentence that follows "今天早上,他吃了很多苹果 (This morning, he ate a lot of apples)" will be able to fill in the correct answer "爱 (love)" or "喜欢 (like)" in the previous sentence.

3.3 Item writing for pseudo-character section

As pseudo-character definitions and classifications have not yet reached a consensus, and as participants in this study are CFL beginners with limited understanding of Chinese characters, the pseudo-characters are thus defined in the current study as non-existent characters that are artificially constructed to look like real Chinese characters by changing orthographic units, such as strokes and radicals.

The following considerations led to the selection of the five characters in this study; 看 (to look), 做 (to do), 爱 (to love), 热 (hot), and 哪 (which) (see Figure 4). As this study focuses on developing an achievement test, the syllabus was strictly followed when developing test items. The characters were taken from the main textbook used by the participants, while the EBCL (European Benchmarking Chinese Language) Proposed List of A1 characters and the vocabulary list in *New Practical Chinese Reader Volume 1* (Liu, 2015) were also consulted to ensure that the characters chosen are essential for CFL beginners, and their parts of speech are consistent.


B.Tick the real Chinese characters, write down the Pinyin and translate the character into English. (0.5*3*5=7.5).

① 做( ) 做( ) 做( ) 做( ) Pinyin:_____ English Translation: _____

② 爱( ) 受( ) 爱( ) 发( ) Pinyin:_____ English Translation: _____

③ 看( ) 香( ) 香( ) 香( ) Pinyin:_____ English Translation: _____

④ 热( ) 热( ) 扶( ) 垫( ) Pinyin:_____ English Translation: _____

⑤ 哪( ) 哪( ) 哪( ) 哪( ) Pinyin:_____ English Translation: _____

Figure 4: The pseudo-character test developed for the current study.

All Chinese characters selected are among the top Chinese characters in the Modern Chinese Character Frequency List, ranked as follows: 看 (76), 做 (246), 爱 (394), 热 (606), and 哪 (652). The construction of the characters – the radicals and strokes – was also taken into consideration. The number of strokes for most Chinese characters ranges from 6 to 13 (Anderson et al., 2013), with the characters selected in this study within this range at nine strokes (哪 and 看), 10 strokes (爱 and 热), and 11 strokes (做). Most modern Chinese characters are semantic–phonetic compound characters consisting of two major components (Anderson et al., 2013). As compound characters offer more possibilities for constructing pseudo-characters, all characters selected in this study are compound characters. Apart from combining two orthographic units that swap positions in Chinese characters to form pseudo characters (e.g., Chen et al., 1996), this study makes further attempts to produce pseudo characters, as shown in Table 1.

Table 1: Creating pseudo-characters for the current study.

Action	Examples	
	Original character	Result
1. Adding or deleting strokes	爱	发
2. Repositioning the position of orthographic units within a character	做	傚
3. Utilising similar-shaped components to replace the originals	热	垫
4. Utilising components with related meaning to replace the originals	哪	哪

Finally, as "recognising a Chinese character generally means decoding both its pronunciation (phonology) and meaning (semantics)" (Zhang et al., 2022, p. 1), this study requires students to not only identify a real character from four options (three of which are pseudo-characters), but also to provide its Pinyin and English meaning.

3.4 Participants

A total of 53 participants took part in the study, 27 males and 26 females. Their average age was 19.39, with a range of 17–22. All participants were enrolled in a beginner's Chinese module in an Irish university (where the researchers were teachers) and were therefore recruited on account of convenience. As this is an introductory-level course open to all students in the university, the majority of participants (62.26 %) were first-year students, while 15 participants (28.30 %) and five participants (9.43 %) were second- and third-year students respectively. In addition, the module was taken by 49.17 % of the participants as an elective, with the remainder taking it as a core or option module closely related to their academic programme.

3.5 Item weighting

In this study, the items are weighed according to their complexity and characteristics, and the overall score is calculated by adding up each questions' marks. The pseudo-character section consists of five Chinese characters containing 15 items in total. For each character, the participants were tested on three separate items of form, phonology, and meaning, with an equal weighting of 0.5 points for each item. The 15 items on the C-test are more challenging than those in the pseudo-character section as they are designed to measure a student's comprehensive understanding of Chinese. Consequently, the C-test is assigned higher marks; one point per item. Since most questions involve Chinese character writing, participants should not lose their entire mark for minor Chinese character errors – this would defeat the purpose of some questions. A detailed description of item weighting can be found in Table 2.

Table 2: Overview of items and item weighting.

Type	Content	Score	*N*	Percentage	Item number
Pseudo-character section	做	1.5	3	20 %	P1 P2 P3
	爱	1.5	3	20 %	P4 P5 P6
	看	1.5	3	20 %	P7 P8 P9
	热	1.5	3	20 %	P10 P11 P12
	哪	1.5	3	20 %	P13 P14 P15

Table 2: (continued)

Type	Content	Score	*N*	Percentage	Item number
C-test section	Phonology	2	2	13.33 %	C1 C14
	Character & Component	3	3	20 %	C3 C7 C12
	Morpheme & Word	4	4	26.66 %	C4 C6 C11 C13
	Grammar	4	4	26.66 %	C2 C5 C8 C15
	Discourse & Cohesion	2	2	13.33 %	C9 C10

3.6 Data collection

In order to conduct the test systematically, the following steps were taken (Figure 5).

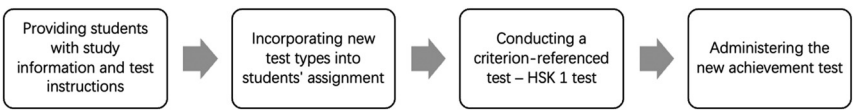


Figure 5: The process of data collection.

Firstly, after ethical approval was received, all participants were informed that the results of the C-test, the pseudo-character section, and the criterion-referenced test would be used anonymously in the current research once the participants granted permission. Secondly, a C-test and a pseudo-character exercise were assigned 2 weeks prior to the test. This gave the participants an opportunity to practise the new exercises and to ask any questions. Although none of the participants expressed confusion regarding the new exercises, a few instructions were later rephrased to ensure clarity.

A sample HSK 1 test was carried out during regular class hours 1 week prior to the new test as a criterion-referenced test. Statistically, a criterion-referenced test only requires a certain number of samples (Zhang, 2016). A sample size of over 32 % can be considered a reasonable for criterion-referenced tests based on previous research (Feng et al., 2020). In the current study, 70 % of the participants (37 out of 53) took the sample HSK 1 test.

A week after the module was completed, the newly developed achievement test – which formed part of participants' final grade – was administered. Detailed instructions were announced by the examiners as well as printed in English on the test paper. To minimise the risk of pseudo-character results being contaminated, this achievement test was divided into two testing phases. In the first phase, a

pseudo-character section with 15 items was to be completed in 5 min. All participants were required to complete the items on a separate test paper and turn it in before proceeding to the second phase. In the second phase, 15 C-test items were included on the test paper.

3.7 Scoring

In terms of scoring, the two researchers who developed the new test also took on the role of raters. As a precaution against bias, both raters marked the tests with student identity completely concealed. To minimise the possibility of unfairness in test paper scoring due to individual differences between raters, one rater was assigned to grade the C-test and a second rater was solely responsible for grading the pseudo-character section. Following the first round of scoring, the test results were reviewed and recalculated by two raters to ensure accuracy.

3.8 Methods of post-test analysis

This study aims to examine whether the newly developed pseudo-character test and C-test can accurately assess student achievement. To address this, it adopts mixed methods to assess descriptive statistics, difficulty, discrimination power, validity, and reliability. SPSS software (version 26) and the online platform SPSSAU (2021) are used for quantitative analysis, while a qualitative analysis is performed to examine the content validity of the new test.

4 Results

4.1 Descriptive statistics

Table 3 shows the descriptive statistics for the new test. The mean score of 13.84 represents a fair level of achievement (61.45 %). In this study, the standard deviation (SD) of 5.5079 indicates that the test seems to be capable of distinguishing between students with different language levels, as the participants' scores cover a relatively wide range.

However, there is a significant gap between high scores and low scores – the lowest score is 2.25 and the highest score is 22.5. Table 3 displays the individual descriptive statistics for the C-test and the pseudo-character section. The mean score of 5.4848 in the pseudo-character section is relatively high since the overall score is

Table 3: Descriptive statistics for each section of the newly designed achievement test.

		Pseudo-character section	C-test section	The achievement test
N	Valid	53	53	53
	Missing	0	0	0
Mean		5.4858	8.3302	13.8160
Minimum		1.25	1	2.25
Maximum		7.5	15	22.5
Std deviation (SD)		0.7071	4.1855	5.5079

only 7.5. Correspondingly, due to the high scores achieved by most participants, the SD of the pseudo-character section is quite low, at 0.7071. In the C-test, the mean score is 8.3302 and the SD is 4.1855, which is more satisfactory as students learn at different rates and a dispersion around a mean is to be expected for an achievement test.

In this study, a normal distribution analysis was conducted as a prerequisite to further analysis. As shown in the Q–Q plot (Figure 6), the distribution of scores appears to be normal. However, to ensure accuracy, a Shapiro-Wilk test was also employed. This resulted in Sig.=0.074 (alpha level=0.05), confirming the null hypothesis that the test results of this study come from a normally distributed population.

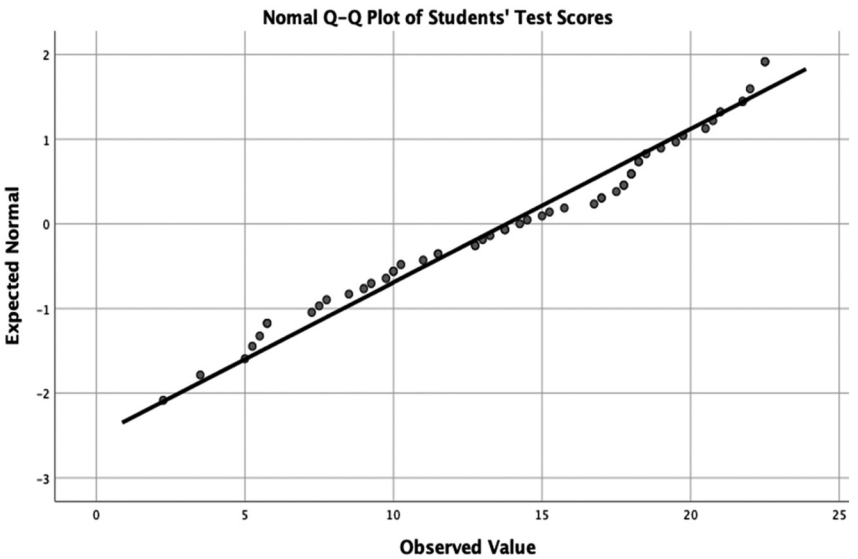


Figure 6: Q–Q plot of scores.

4.2 Item analysis

Item analysis examines student responses to individual test items in order to assess the quality of those items and of the test as a whole (Gajjar et al., 2014). Testers can use item analysis to determine which items should be adopted, revised, or discarded. Difficulty (henceforth *p*-value) and discrimination power (henceforth DP) are the two main factors assessed in the current study.

4.2.1 Difficulty (*p*-value)

The items on the C-test were scored on a scale of 0–1, while the pseudo-character marks ranged from zero to the full 0.5 marks. Therefore, *p*-value formulas (see Figure 7) for a constructed-response or performance item are employed in this study (Brookhart & Nitko, 2019; Zhang, 2016).

$$\text{Difficulty index (}p\text{-value)} = \frac{\text{Average score - minimum possible score}}{\text{Maximum possible score - minimum possible score}} = \frac{\text{Average score}}{\text{Range of full marks}}$$

Figure 7: Formula of difficulty index.

The range and recommendation of *p*-value used in this study are commonly used by scholars for educational domains (Johari et al., 2011) (see Table 4). As shown, the majority of the 21 items (70 %) are of acceptable difficulty, with a *p*-value of 30–80 %. Only one item in the C-test falls below 0.3, while eight items – one from the C-test and seven from the pseudo-character section – are above 0.80. The median difficulty level of the newly designed achievement test is 0.64, slightly more difficult than expected (ideally between 0.5 and 0.6 [Escudero et al., 2000]).

Table 4: Level of difficulty (*p*-value).

Level of difficulty	*p*-Value range	Question number	*N*	%
Very difficult	<0.30	C15	1	3.33 %
Moderate	0.30–0.80	C1/C3/C4/C5/C6/C7/C8/C9/C10/ C11/C12/C13/C14 P2/P5/P6/P8/P11/P12/ P14/P15	21	70.00 %
Very easy	>0.8	C2 P1/P3/P4/P7/P9/P10/P13	8	26.67 %

[a]C refers to the items in the C-test section. [b]P refers to the items in the pseudo-character section.

4.2.2 Discrimination power (DP)

DP is the ability of an item to discriminate between higher- and lower-achieving students (Rezigalla, 2022). A more reliable assessment will be composed of tasks with high positive DP indices (Brookhart & Nitko, 2019). DP indices can reach a maximum of 1.00. The range and interpretation of the DP indices used in the present study (see Table 5) were adopted in a number of previous studies (e.g., Aljehani et al., 2020; Bhattacherjee et al., 2022).

Table 5: The DP of each item and its interpretation.

DP range	Quality	Items		Item details	Recommendation
		N	%		
≥0.4	Very good item	24	80.00	C1/C3/C4/C6/C7/C8/C9/C10/C11/C12/C13/C14/C15 P2/P3/P4/P5/P6/P8/P9/P10/P11/P12/P14	Keep
0.30–0.39	Good item	1	3.33	C2	Keep
0.20–0.29	Fair item	3	10.00	P15/P7 C5	Keep
<0.20	Marginal item	2	6.67	P13/P1	Revise/discard
Negative	Worst/defective item	0	0.00		Discard

[a]C refers to the items in the C-test section. [b]P refers to the items in the pseudo-character section.

Table 5 reveals that 24 (80 %) of the 30 items have excellent DP, 3 % are good items, 10 % are fair items, and 6.67 % are relatively poor. The highest DP was item C13 (0.784) and the lowest DP was item P13 (0.197). Items P13 and P1 need to be revised or discarded, as their DP value is lower than 0.2.

4.3 Validity

4.3.1 Criterion (concurrent) validity

Criterion behaviour should occur simultaneously or nearly simultaneously with the administration of the test (Bachman, 1990). Based on Cronbach's Alpha, the HSK 1 test has a good degree of reliability, with indices of 0.85–0.95 (Luo et al., 2011). Therefore, the sample HSK 1 test was conducted 1 week before the new achievement test was administered. Furthermore, Pinyin was removed from this sample HSK 1 test in advance to ensure that it did not significantly reduce the difficulty of the test.

According to the results (see Table 6), the newly designed achievement test shows significant positive correlation with the sample HSK 1 test – in relation to overall scores (0.848), listening scores (0.719), and reading scores (0.804). The score of the C-test also has a high correlation coefficient with the overall scores (0.809), the reading scores (0.864), and the listening scores (0.748) of the sample HSK 1 test. In comparison, the correlation between the pseudo-character section and the sample HSK 1 test was less favourable, ranging from 0.495 to 0.636. Possibly, this is due to the low difficulty or an insufficient number of items in the pseudo-character section.

Table 6: Correlation between the newly designed test and sample HSK1 test.

	The achievement test	Pseudo-character section	C-test section	Sample HSK 1 test (reading)	Sample HSK 1 test (listening)	Sample HSK 1 test (overall)
The achievement test	1	0.852**	0.983**	0.804**	0.719**	0.848**
Pseudo-character section	0.852**	1	0.740**	0.633**	0.495**	0.636**
C-test section	0.983**	0.740**	1	0.809**	0.748**	0.864**
Sample HSK 1 test (reading)	0.804**	0.633**	0.809**	1	0.634**	0.938**
Sample HSK 1 test (listening)	0.719**	0.495**	0.748**	0.634**	1	0.863**
Sample HSK 1 test (overall)	0.848**	0.636**	0.864**	0.938**	0.863**	1

**Correlation is significant at the 0.01 level (2-tailed).

4.3.2 Content validity

In this study, four experienced language teachers with a Master's in Teaching Chinese to Speakers of Other Languages were recruited as experts. In addition to having at least 5 years teaching experience with CFL beginners, they are currently teaching Chinese in China, the UK, Ireland, and the US. The content validity index (CVI), as proposed by previous scholars (Elangovan & Sundaravel, 2021; Souza et al., 2017) was adopted in this study. The experts were required to examine each item on a four-point scale: 1=non-equivalent item; 2=the item needs to be extensively revised; 3=equivalent item, needs minor adjustments; and 4=totally equivalent item.

The I-CVI refers to the content validity of each individual item whereas the S-CVI/Ave refers to the average of the I-CVIs for all items on the scale (Polit & Beck, 2006). As a whole, 25 out of 30 items in this study have an I-CVI=1.00, four items are 0.75, and one item is 0.5. As advocated by Lynn (1986), when there are five or fewer experts, a scale that is considered excellent in content validity should be composed of items with I-CVI=1.00 and with S-CVI/Ave of 0.90 or higher. Therefore, the majority of items (83.33 %) were considered relevant, except for five items; C5, C12, and C13 in the C-test, and P1 and P13 in the pseudo-character section. Furthermore, this study also has an acceptable S-CVI/Ave of 0.95.

The judges explained their decisions regarding the items with low I-CVIs. In particular, C5 and C13 were rated 1 or 2 since the proposed answers "也" and "天气" may not be the only options. C12 has a relatively low rate with 0.75, as ideally "木" needs to be presented in the form of a radical, with a shorter fourth stroke "⺀". Manually formed pseudo-characters "哪" and "做" in P1 and P13 have an unnatural font, which may lead to guessing behaviour.

Content validity concerns not only linguistic skills, but also the setting, which consists of administration, format, and time constraints (Bachman, 1990). The experts were thus also interviewed with 15 questions (see Appendix I) regarding the test settings. A number of suggestions were raised by the experts to enhance the pseudo-character test, including adding items, increasing its weighting and updating the technology of generating pseudo-characters. They also proposed using the test in classroom exercises instead of assessments to avoid causing anxiety in students. Regarding the C-test, the experts highlighted the requirement for each answer to be unique. Overall, the new achievement test was commended for its innovativeness and comprehensiveness.

4.3.3 Construct validity

According to Bachman (1982), if the test results match the model constructed when the questions were designed, this demonstrates that the test has some structural validity. Confirmatory factor analysis (CFA), widely regarded as a powerful tool for establishing evidence supporting construct validity of language tests, is adopted in this study. First, the model for the C-test is initially based on the categorisation of language competence, while the pseudo-character test model is based on the selected characters (see Table 2). In addition, according to previous research (Bachman, 1982; Fan et al., 2014), a higher-factor model incorporating one superordinate factor and a series of subordinate factors upon which a specified subgroup of items exist provide a satisfactory fit for language test data. As a result, higher-factor hypothetical models for the C-test and the pseudo-character section have been developed and are schematically illustrated in Figures 8 and 9.

To determine the models' fit in the present study, two criteria are employed. Criteria one consists of selected global fit indices: (1) the comparative fit index (CFI); (2) the minimum discrepancy per degree of freedom (CMIN/DF); (3) the root mean square error of approximation (RMSEA); and (4) the standardised root mean square residual (SRMR). For CFI, values greater than 0.9 represent a good fit (Awang, 2012). Kline (1998) suggests an acceptable fit when CMIN/DF is <3. For RMSEA and SRMR, values of 0.05 or lower are indicative of a good and excellent fit (Burns & Patterson, 2000). A second criterion is the value of two psychometric property indicators, the composite reliability (CR) and the average variance extracted (AVE). If the minimum value of CR exceeds 0.60, AVE values above 0.40 are acceptable (Fornell & Larcker, 1981).

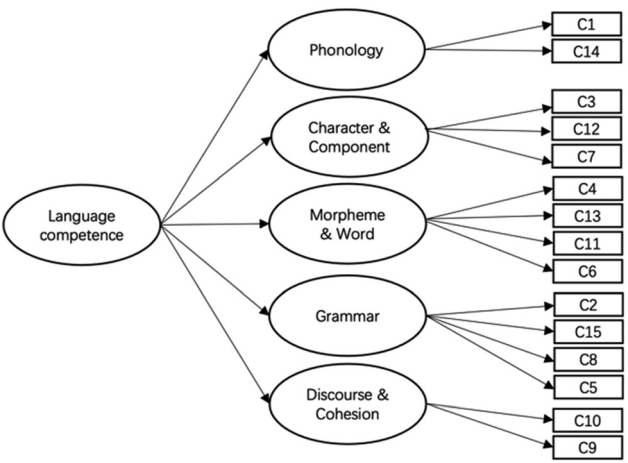


Figure 8: Hypothetical model for C-test.

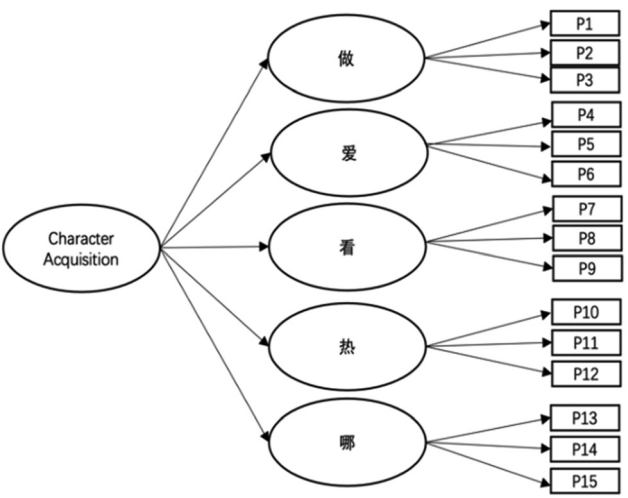


Figure 9: Hypothetical model for pseudo-character section.

The online platform SPSSAU (2021) was employed to assess the fits of the models in this study. According to the Table 7, the C-test has a good overall model fit with χ^2/df=1.153, CFI=0.996, SRMR=0.073, and RMSEA=0.054. Pseudo-character test results were also satisfactory with χ^2/df=1.370, CFI=0.914, SRMR=0.090 and RMSEA=0.084.

Table 7: Fit indices for the models.

	CMIN/DF (χ^2/df)	CFI	SRMR	RMSEA
C-test	1.153	0.996	0.073	0.054
Pseudo-character section	1.370	0.914	0.090	0.084
Reference values	<3	>0.9	<0.05	<0.05

Regarding the second criterion, satisfactory AVE values were obtained for the pseudo-character section; however, a slightly unsatisfactory CR value of 0.584 was obtained for the character "哪". For the C-test, the results were more complicated. For the factors Phonology, Character & Component, and Morpheme & Word, the results were satisfactory. For the factor Discourse & Cohesion, although a fairly satisfactory value of 0.422 was obtained for AVE, a slightly unsatisfactory value of 0.587 was produced for CR. A critical finding was that the AVE value for the factor Grammar was 0.113 and the CR value was 0.294, suggesting that the four variables under the grammar factor were not as well converged (see Table 8).

Table 8: AVE and CR results for the hypothetical models.

C-test section			Pseudo-character section		
Factor	**AVE**	**CR**	**Factor**	**AVE**	**CR**
Phonology	0.494	0.659	Character 做	0.481	0.703
Character & Component	0.423	0.683	Character 爱	0.666	0.854
Morpheme & Word	0.558	0.834	Character 看	0.496	0.732
Grammar	0.113	0.294	Character 热	0.617	0.806
Discourse & Cohesion	0.422	0.587	Character 哪	0.458	0.584

Reference values: CR > 0.60, AVE > 0.40.

4.4 Reliability

Besides referring to the quality of the test scores, reliability also refers to their consistency. Internal consistency reliability, which focuses on the consistency of a test's internal elements, is assessed in the current study. In theory, Cronbach's Alpha

can be applied when test items are weighted and vary from 0 to 1 (Jin & Qi, 2018). A higher Cronbach's Alpha indicates greater relevance between items and greater internal consistency of reliability.

Table 9 reveals that the overall reliability of the new achievement test is good, with a Cronbach's Alpha of 0.907. This indicates that the new achievement test is reliable to a degree of 90.7 %. Two of the subscales, the C-test and the pseudo-character section, have Cronbach's Alpha coefficients of 0.875 and 0.840, respectively, which also indicates high levels of reliability.

Table 9: Internal consistent reliability.

Scale	Cronbach's Alpha	*N*
The achievement test	0.907	30
C-test section	0.875	15
Pseudo-character section	0.840	15

5 Discussion

The assessment types used in Chinese language tests should not be limited to those already in use. Researchers and teachers of Chinese language should take inspiration from successful tests in other languages and incorporate new question types into Chinese language teaching and testing through reasonable adaption. Taking inspiration from a booming online English test, this study attempted to develop an achievement test containing two new question types: the C-test and the pseudo-character test, for use with beginner-level Chinese learners. Statistically, the C-test and pseudo-character test appear to be valid and reliable in the context of CFL.

5.1 The development and accuracy of the C-test and pseudo character test in the CFL classroom

As for the overall design of the achievement test, it differs significantly from the HSK 1 test in which the knowledge to be tested is randomly distributed over 40 questions that are not contextually relevant. In contrast, the newly developed achievement test is clearly composed of two main types of question, with the C-test assessing students' overall competence in phonology, Chinese characters, vocabulary, grammar, and discourse, and the pseudo-character section focusing on beginners' understanding of Chinese characters. In the redesigned C-test, the concept of "damage" has been incorporated and items are generated with damaged characters, words, and

phrases. In addition, it tests Pinyin knowledge, which is rarely included in CFL assessments. The detailed instructions and broad range of content areas are among the innovations of the newly designed C-test, while all judges noted innovation to be one of the greatest strengths of the pseudo-character section. The attempts to select, define, and produce pseudo characters in this study can serve as inspiration to CFL teachers. Certainly, some drawbacks of these two question types were identified during the design process. For instance, for beginner students, learning Chinese characters is already a challenging task, and the pseudo-character section may further complicate the learning process and result in student anxiety. In addition, since the design of a programme to automatically generate pseudo-characters also requires the assistance of professionals, it is not currently possible to generate pseudo-characters in a time-saving manner using a single font.

Looking closely at the post-test analysis, a normal distribution with acceptable mean scores and SD provides evidence that the test can differentiate language levels to some extent. Regarding the item analysis, the overall difficulty of the test paper is slightly higher than expected. In particular, the pseudo-character section seems to be too easy, with seven questions classified as “very easy”. In contrast, the difficulty of C-test items is fairly good, with only one question ranked “very easy” and one “very difficult”. Approximately 80 % of the DP items are considered very good items, and none should be eliminated completely, since none have negative values. Satisfactory Cronbach’s Alpha values also indicate that the newly designed achievement test is highly consistent and reliable.

At the same time, the three major and complementary types of validity tests also reveal promising and interesting results. Firstly, having significant correlations with the sample HSK 1 test provides robust evidence for the criterion-related (concurrent) validity of the newly designed achievement test. Both the C-test and the pseudo-character section show significant correlations with the listening section and the reading section of the criterion. In other words, after further exploration, these two novel tests may have the potential to serve as a proficiency test for evaluating students’ overall language competence. With regard to construct validity, the CFA provides evidence that both the C-test and the pseudo-character section fit the hypothetical higher-order model fairly well, indicating that the newly developed achievement test reflects the original intentions of the test designers. Due to the fact that the four variables in the grammar factor in the C-test were not sufficiently converged, improvements can still be made.

The current paper has therefore addressed the research questions in both demonstrating how a C-test and pseudo-character test can be developed in the context of CFL, as well as identifying the accuracy of such in assessing students’ learning outcomes.

5.2 Implications for online testing

The digital-age learning environment shifts away from preparing language learners for traditional written tests to creating innovative, comprehensive assessments that reflect their progress and achievement. The newly developed C-test and pseudo-character test also have a high potential to be adopted in online assessments.

The artificially created pseudo-characters could be displayed as non-editable and non-Googleable images, thereby further reducing cheating and making OCR impossible. Also, although typing is not an appropriate method of assessing the composition of characters (Guan et al., 2011), the pseudo-character test mitigates this shortcoming by examining students' knowledge of components and strokes. In addition, the pseudo-character test is not exclusively limited to formative assessments but can also be utilised as a summative assessment, a homework assignment, and an interactive online classroom activity or game. For example, by using a smartphone or tablet, students can identify the real characters among the pseudo-characters or even provide short answers detailing incorrect components or strokes in online polling applications, such as *Kahoot!*, *Socrative*, and *Plickers*. Pseudo-character tests also provide inspiration for technological specialists, for instance, programming techniques can be utilised to generate pseudo-characters automatically.

The C-test is also ideal for online testing with the potential to be developed as a computer-adaptive test (CAT). Another advantage of the C-test is the definite answers. With objective and exclusive answers, the C-test can be scored without much subjective judgement and can therefore be administered online with an auto-scoring system. Moreover, technologies such as machine learning and Natural Language Processing (NLP) are widely employed to automatically create cloze-test items as well as building large-scale cloze datasets (Correia et al., 2012; Xie et al., 2017). In everyday settings, the C-test can also be applied to online CFL classrooms. Using collaborative digital whiteboard platforms such as *Padlet* or *Miro*, students can complete C-test exercises in groups. The design concept of tile-matching video games, such as *Tetris*, can be employed in C-test exercises, whereby the sentence changes colour once all the blanks in the sentence have been correctly filled in by the students. Moreover, the C-test can also be designed as a simple matching-up or drop-down online practice.

Overall, the newly designed C-test and pseudo-character section have great potential and advantages to be adapted for the online space. The next step of the research will be to introduce them to online platforms, thus making them available for both the at-distance and face-to-face CFL classroom.

6 Conclusions

The study adapts two different types of questions – the C-test and pseudo-character test – for beginner CFL learners. The researchers followed the key steps of assessment development and the statistical results provide strong support for the reliability and validity of the newly developed tests, demonstrating that the two new question types can be successfully incorporated into the Chinese test with appropriate modifications.

This study is of great significance in at least four aspects: (1) it provides an innovative and achievable approach on how to implement C-tests and pseudo-character tests in the CFL context; (2) it provides a valid and reliable test to aid other Chinese language educators with the same needs; (3) it provides insight into a comprehensive, reliable, and practical methodology for developing Chinese language tests; and (4) it suggests how C-tests and pseudo-character tests may be applied to the digital age.

The study is not without limitations. Due to the limited number of judges who participated in testing the content validity, the results obtained should be interpreted with caution. Additionally, a small sample size may have negatively affected the stability and accuracy of CFA analysis, while test-retest reliability statistics are missing in this study, as the researchers were unable to administer a second test to all participants after the university semester had ended.

A number of potential directions for future research could be considered. Initially, it would be beneficial to examine whether listening and speaking tests could be incorporated into the new test. Students could be required to respond orally to questions regarding the text on the current C-test, for example. Moreover, the use of pseudo-character tests and C-tests in Chinese language proficiency and placement assessments requires further exploration. It is also possible to further develop the pseudo-characters questions with more context provided. For example, providing “做饭” so that test takers understand the contextual meaning of “做”, and “看书” as additional information for “看”. Finally, having the newly designed C-test and pseudo-character test applied in online Chinese assessments that can be administered remotely and scored automatically would be an even greater achievement.

Research funding: This work is supported by Center for Language Education and Cooperation [Grant No. 21YH95D]. Title of the funding: 2021 International Chinese Language Education Research Project (2021 年国际中文教育研究课题).

Appendix I

Questions Given to the Judges for Assessing Content Validity

Q1	Is the overall test (including the pseudo-character test and the C-test) appropriate for use as an achievement test?
Q2	Would you consider this test to be at the appropriate level for the students?
Q3	Would you consider assigning the C-test and a pseudo-character exercise 2 weeks prior to the test to be appropriate?
Q4	Would you consider the instructions for the pseudo-character test are clear and easy to follow?
Q5	Would you consider the instructions for the C-test are clear and easy to follow?
Q6	Would you consider the weighting assigned to each item in the pseudo-character test reasonable?
Q7	Would you consider the weighting assigned to each item in the C-test reasonable?
Q8	Would you consider 5 min to be a suitable amount of time for the pseudo-character test?
Q9	Would you consider it appropriate to require participates complete the pseudo-character test and submit it before the C-test begins?
Q10	Would you consider it reasonable to assign one rater to grade the C-test and a second rater to grade the pseudo-character test?
Q11	Would you consider the order and presentation of the pseudo-character test reasonable?
Q12	Would you consider the order and presentation of the C-test reasonable?
Q13	What are the advantages of the overall test?
Q14	What are the disadvantages and problems of the overall test?
Q15	Other comments

References

Alderson, C. J., Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, *13*(2), 219–227.

Aljehani, D. K., Pullishery, F., Osman, O. A. E., & Abuzenada, B. M. (2020). Relationship of text length of multiple-choice questions on item psychometric properties – a retrospective study. *Saudi Journal for Health Sciences*, *9*(2), 84.

Alsied, F. M., & Pathan, M. M. (2013). The use of computer technology in EFL classroom: Advantages and implications. *International Journal of English Language & Translation Studies*, *1*(1), 44–51.

Anderson, R. C., Ku, Y. M., Li, W., Chen, X., Wu, X., & Shu, H. (2013). Learning to see the patterns in Chinese characters. *Scientific Studies of Reading*, *17*(1), 41–56.

Awang, Z. (2012). *Structural equation modeling using amos graphic*. Amsterdam University Press.

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, *16*(1), 61–70.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535–556.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
Barrette, C. (2004). An analysis of foreign language achievement test drafts. *Foreign Language Annals*, *37*(1), 58–70.
Bhattacherjee, S., Mukherjee, A., Bhandari, K., & Rout, A. J. (2022). Evaluation of multiple-choice questions by item analysis, from an online internal assessment of 6th semester medical students in a rural medical college, West Bengal. *Indian Journal of Community Medicine*, *47*(1), 92–95.
Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). Upper Saddle River, N.J Pearson Merrill Prentice Hall.
Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. Longman.
Burns, D., Dagnall, N., & Holt, M. (2020). Assessing the impact of the COVID-19 pandemic on student wellbeing at universities in the United Kingdom: A conceptual analysis. *Frontiers in Education*, *5*(582882), 1–10.
Burns, G. L., & Patterson, D. R. (2000). Factor structure of the Eyberg child behavior inventory: A parent rating scale of oppositional defiant behavior toward adults, inattentive behavior, and conduct problem behavior. *Journal of Clinical Child Psychology*, *29*(4), 569–577.
Cardwell, R., LaFlair, G. T., & Settles, B. (2022). Duolingo English test: Technical manual. *Duolingo Research Report*. https://duolingo-papers.s3.amazonaws.com/other/det-technical-manual-current.pdf
Chang, L. Y., Tseng, C. C., Perfetti, C. A., & Chen, H. C. (2022). Development and validation of a Chinese pseudo-word/non-character producing system. *Behavior Research Methods*, *54*(2), 632–648.
Chen, Y. P., Allport, D. A., & Marshall, J. C. (1996). What are the functional orthographic units in Chinese word recognition: The stroke or the stroke pattern? *The Quarterly Journal of Experimental Psychology*, *49A*, 1024–1043.
China Education Center. (2023). HSK test – new HSK test 3.0. https://www.chinaeducenter.com/en/hsk/newhsk3.php#:~:text=The%20new%20HSK%20test%203.0%20uses%20Four%2Ddimension%20Benchmarks%20include,specific%20requirements%20for%20each%20dimension
Chinese Testing International. (2022a). 考试规则-汉语考试服务网. www.chinesetest.cn/gotestlaw.do#
Chinese Testing International. (2022b). 考试介绍-汉语考试服务网. www.chinesetest.cn/gosign.do?id=1&lid=0
Chung, K. K., Tong, X., & McBride-Chang, C. (2012). Evidence for a deficit in orthographic structure processing in Chinese developmental dyslexia: An event-related potential study. *Brain Research*, *1472*, 20–31.
Correia, R., Baptista, J., Eskenaz, M., & Mamede, N. (2012). Automatic generation of cloze question stems. In H. Caseli, A. Villavicencio, A. Teixeira, & F. Perdigão (Eds.), *Computational Processing of the Portuguese Language* (pp. 168–178). Springer.
Daniels, L. M., Goegan, L. D., & Parker, P. C. (2021). The impact of COVID-19 triggered changes to instruction and assessment on university students' self-reported motivation, engagement and perceptions. *Social Psychology of Education*, *24*, 299–318.
Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of Language Testing* (Vol. 7). Cambridge University Press.
DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems*. University of Hawaii.
Dobrić, N. (2018). Conceptualization of validity in educational testing – historical discussion and contemporary consensus. *AAA: Arbeiten aus Anglistik und Amerikanistik*, *43*(1), 3–26.
Duolingo. (2023). Duolingo English test. https://englishtest.duolingo.com/applicants
Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-test. *Language Testing*, *23*(3), 290–325.
Elangovan, N., & Sundaravel, E. (2021). Method of preparing a document for survey instrument validation by experts. *MethodsX*, *8*, 101326.

Escudero, E. B., Reyna, N. L., & Morales, M. R. (2000). The level of difficulty and discrimination power of the Basic Knowledge and Skills Examination (EXHCOBA). *Revista Electrónica de Investigación Educativa*, *2*(1), 2.

Fan, J., Ji, P., & Yu, L. (2014). Another perspective on language test validation: Investigating the factor structure of language tests. *Foreign Language Learning Theory and Practice*, *4*(2), 34–40.

Feng, L., Feng, H., Bai, S., & Wu, J. (2020). 汉语二语水平快速测试的试卷研发分析—基于等距离完形填空的研究 [An analysis of a proficiency test for CSL (Chinese as second language) based on fixed-ratio cloze questions]. *Applied Linguistics*, (3), 69–79.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine*, *39*(1), 17.

Grotjahn, R. (2002). Konstruktion und Einsatz von C-tests: Ein Leitfaden für die Praxis [Construction and use of C-test: A guide for practical implementation]. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* [The C-test. Theoretical foundations and practical applications] (pp. 211–225). AKS.

Guan, C., Liu, Y., Chan, D., Ye, F., & Perfetti, C. A. (2011). Writing strengthens orthography and alphabetic-coding strengthens phonology in learning to read Chinese. *Journal of Educational Psychology*, *103*(3), 509–522.

Harsch, C., & Hartig, J. (2016). Comparing C-tests and Yes-No vocabulary size tests as predictors of receptive language skills. *Language Testing*, *33*(4), 555–575.

Huyen, N. T. (2022). The effects of using online applications to teach vocabulary to English learners of HUFI in Ho Chi Minh City. *International Journal of TESOL & Education*, *2*(3), 32–42.

Jiang, L. (2014). *HSK standard course 1 Vol. 1 – Textbook* (1st ed.). Beijing Language & Culture University Press.

Jin, L., Xu, Y., Deifell, E., & Angus, K. (2021). Emergency remote language teaching and U.S.-based college-level world language educators' intention to adopt online teaching in postpandemic times. *The Modern Language Journal*, *105*(2), 412–434.

Jin, Y., & Qi, X. (2018). The SPSS-based analysis of reading comprehension – take grade eight English mid-term test, for example. *Journal of Language Teaching and Research*, *9*(5), 939–945.

Johari, J., Sahari, J., Abd Wahab, D., Abdullah, S., Abdullah, S., Omar, M. Z., & Muhamad, N. (2011). Difficulty index of examinations and their relation to the achievement of programme outcomes. *Procedia-Social and Behavioral Sciences*, *18*, 71–80.

Khansir, A. A., & Pakdel, F. (2018). Place of error correction in English language teaching. *Educational Process: International Journal*, *7*(3), 189–199.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47–84.

Kline, R. B. (1998). *Structural Equation Modeling*. Guilford.

Kuo, L. J., Kim, T. J., Yang, X., Li, H., Liu, Y., Wang, H., Park, J. H., & Li, Y. (2015). Acquisition of Chinese characters: The effects of character properties and individual differences among second language learners. *Frontiers in Psychology*, *6*, 986.

Liu, X. (2015). *New practical Chinese reader Vol. 1 – Textbook* (3rd ed.). Beijing Language & Culture University Press.

Liu, Y., Wang, M., & Perfetti, C. A. (2007). Threshold-style processing of Chinese characters for adult second-language learners. *Memory & Cognition*, *35*(3), 471–480.

Luo, M., Zhang, J., Xie, O., Huang, H., Xie, N., & Li, Y. (2011). 新汉语水平考试 (HSK) 质量报告 [New Chinese proficiency test (HSK) quality report]. *China Examinations*, (10), 3–7. https://doi.org/10.19360/j.cnki.11-3303/g4.2011.10.001

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*, 382–385.
McNaughton, W. (2005). *Reading & writing Chinese simplified character edition: A comprehensive guide to the Chinese writing system* (3rd ed.). Tuttle Publishing.
Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.
Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Insights into Non-Native Vocabulary Teaching and Learning* (Vol. 52, pp. 83–98). Multilingual Matters.
Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Newbury House Publishers.
Oller, J. W. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *The Modern Language Journal*, *56*(3), 151–158.
Oller, J. W. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning*, *23*(1), 105–118.
Osborne, C., Zhang, Q., & Zhang, G. X. (2018). Which is more effective in introducing Chinese characters? An investigative study of four methods used to teach CFL beginners. *Language Learning Journal*, *48*, 1–17.
Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489–509.
Peng, Y., Yan, W., & Cheng, L. (2021). Hanyu Shuiping Kaoshi (HSK): A multi-level, multi-purpose proficiency test. *Language Testing*, *38*(2), 326–337.
Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497.
Raatz, U., & Klein-Braley, C. (1981). The C-test: A modification of the cloze procedure. In C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing* (Vol. 7, pp. 113–138). University of Essex.
Ran, Q., & Yu, S. (2019). 汉语语音偏误的特点与模式——基于25种母语背景学习者的偏误条目数据的分析 [Characteristics and patterns of Chinese speech errors – analysis and explorations into the error item dataset from 25 L1 backgrounds]. *Chinese Teaching in the World*, *3*, 417–432.
Rezigalla, A. A. (2022). Item Analysis: Concept and Application. In *Medical Education for the 21st Century* (p. 105). IntechOpen.
Shi, Z. Y. (2000). 外国留学生字形书写偏误分析 [Error analysis of Chinese characters written by foreign learners]. *Chinese Language Learning*, *2*, 38–41.
Souza, A. C. D., Alexandre, N. M. C., Guirardello, E. D. B., & Alexandre, N. M. C. (2017). Psychometric properties in instruments evaluation of reliability and validity. *Epidemiologia e servicos de saude*, *26*, 649–659.
Sze, W. P., Rickard Liow, S. J., & Yap, M. J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, *46*, 263–273.
Tavakol, M., & Dennick, R. (2011). Post-examination analysis of objective tests. *Medical Teacher*, *33*(6), 447–458.
Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433.
VanPatten, B., Trego, D., & Hopkins, W. P. (2015). In-class vs. online testing in university-level language courses: A research report. *Foreign Language Annals*, *48*, 659–668.
Wagner, E. (2020). Duolingo English test, revised version july 2019. *Language Assessment Quarterly*, *17*(3), 300–315.
Wang, D., & East, M. (2020). Constructing an emergency Chinese curriculum during the pandemic: A New Zealand experience. *International Journal of Chinese Language Teaching*, *1*(1), 1–19.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Wu, Q., Hong, W., & Deng, S. (2017). 汉字认读在汉语二语者入学分班测试中的应用—建构简易汉语能力鉴别指标的实证研究 [Application of Chinese character identification in placement tests for CSL learners: An empirical study of constructing simple Chinese proficiency indicators]. *Chinese Teaching in the World*, *31*, 395–411.

Xie, Q., Lai, G., Dai, Z., & Hovy, E. (2017). Large-scale cloze test dataset created by teachers. *arXiv preprint arXiv:1711.03225*.

Ye, J. (2003). 《对外汉语教学语音大纲》初探 [A Preliminary study of the phonetic syllabus for teaching Chinese as a Foreign Language]. *Journal of Yunnan Normal University (Teaching and Research on Chinese as a Foreign Language Edition)*, *1*(4), 62–66.

Zhang, D. (2017). Developments in research on testing Chinese as a second language. In D. Zhang, & C. H. Lin (Eds.), *Chinese as a Second Language Assessment* (pp. 67–87). Springer.

Zhang, H., Kim, S.-A., & Zhang, X. (2022). A comparative study of three measurement methods of Chinese character recognition for L2 Chinese learners. *Frontiers in Psychology*, *13*, 753913.

Zhang, K. (2016). 《语言测试概论：几个问题》 [Introduction to language testing: A few questions]. Beijing Language and Culture University Press.

Zhang, L., Han, H., Zhong, M., & Sun, L. J. (2005). 传统完形填空与 C-试题的效度对比研究 [A comparative study on the validity of traditional Cloze test and C-test]. *Journal of Xinjiang Education Institute*, *21*(3), 63–66.

Zhang, Q. (2020). Narrative inquiry into online teaching of Chinese characters during the pandemic. *International Journal of Chinese Language Teaching*, *1*(1), 20–34.

Zhou, J. (2007). 汉字教学理论与方法 [Theory and methodology of teaching Chinese characters]. Peking University Press.

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures and some correlates of word and non-word recognition. *Intelligence*, *1*(1), 5–13.

Bionotes

Xuan Yang
University College Dublin (UCD) Confucius Institute for Ireland, Dublin, Ireland / Renmin University of China, Beijing, China
xuan.yang@ucd.ie
cucyangxuan@gmail.com
https://orcid.org/0000-0002-1189-7066

Xuan Yang is a full-time Chinese teacher in the international office at Renmin University of China. Since 2017, she has been dispatched by Renmin University of China and the Center for Language Education and Cooperation to work for Confucius Institute at University College Dublin (UCD). She was awarded her bachelor's degree in Teaching Chinese as a Foreign Language from Communication University of China in 2013, and the master's degree in International Comparative Education from Stockholm University in 2015. She currently teaches Chinese language modules at both elementary and intermediate levels in UCD. Her research interests include the compilation of teaching materials, Chinese language testing, Chinese characters teaching, and textbook analysis. Currently, she also takes charge of the Chinese language and cultural evening courses in UCD Confucius Institute.

Caitríona Osborne
University College Dublin, Dublin, Ireland
caitriona.osborne@ucd.ie

Caitríona Osborne is an Assistant Professor in the Irish Institute for Chinese Studies at University College Dublin. She received her BA in Applied Linguistics, MA in Translation Studies, and PhD in Applied Linguistics from Dublin City University. She currently teaches Chinese language and Chinese Teaching Methodology modules associated with the Professional Diploma and Master's in Teaching Chinese Language and Culture in UCD. She has also previously taught Chinese Culture and Translation modules to undergraduate students. Dr Osborne is currently a committee member of the Irish Association for Applied Linguistics and the Irish Association of Chinese Teaching. Her research interests include: Chinese language; Chinese language education; Chinese language pedagogy; Innovative methods for teaching Chinese characters to beginner learners; Translation (Chinese to English). In 2023, Dr Osborne was awarded a UCD College-Level Teaching and Learning award for the year 2021–2022.