Abstract

The study aims at establishing an efficient, valid, and computerized Chinese character recognition assessment in order to meet the increasing needs for online tests in distance education of Chinese. We constructed a Web-based Chinese Character Recognition Assessment system, compiling a test for Chinese Character Recognition Test which included 120 characters from different levels of character frequency in this system. The duration of the assessment was eight minutes. The participants were asked to report the pronunciations of each character that were shown on the screen one at a time. One-week test-retest reliability showed a strong positive correlation, and Cronbach’s Alpha and the split-half reliability demonstrated that this assessment items have good internal consistency. Moreover, a criterion-related validity estimator revealed that the scores of this assessment had significantly positive correlations with three criterions: Graded Chinese Character Recognition Test (Huang, 2001), Tests of Chinese Listening, Speaking, Reading, and Writing (Extension school of Continuing Education of National Taiwan Normal University, 2010), and the three groups of different levels of CEF R (below A1, A1, A2 and above). The results suggest that this online character recognition assessment is valid to evaluate overall Chinese language skills. Moreover, the assessment in this study contains two scoring methods: Tone scoring and non-tone scoring. The results indicate that the tone scoring method already possesses the function of distinguishing the level of students’ Chinese competence (because the score for the character recognition test has a significantly positive correlation with the scores for listening, speaking, reading, and writing tests), and if the non-tone scoring method is adopted (i.e. taking no account of the correctness of tone), the discrimination of items and the validity of testing scores would be enhanced. Theoretically, this study supports that character-level knowledge is one of crucial components of Chinese language knowledge. Practically, this online character recognition assessment is an efficient and valid tool to evaluate the general Chinese ability of those who take Chinese as a second/foreign language.

Keywords: Chinese as second/foreign language, Chinese Character Recognition Assessment, Chinese Language Testing System
「中文線上識字測驗」發展及其在華語遠距教學上的應用

曾千芝 1  張瓅勻 2  張雨霖 3  陳學志 4

摘要

本研究之主要目的為發展一套以中文為第二語言或外語學習者的線上識字測驗，以應用於華語遠距教學上。本測驗包含 120 個字，其是從不同頻率層次下隨機選出。本研究有兩個主要發現：

第一，在信效度研究方面，結果顯示 Cronbach’s Alpha 與折半信度相當良好，間隔一週之再測信度也具有高度相關。在效度研究結果則發現，本測驗具效標關聯效度（效標有三項：「中文認字測驗」、「華語聽說讀寫測驗」，以及不同 CEFR 等級的參與者），支持線上識字測驗分數作為評估整體華語能力指標的有效性。

第二，本測驗包含本研究識字測驗包含兩種計分：完全計分（該字聲韻調全對則得分）與部分計分（該字聲韻全對即得分）。結果發現若採用的是部分計分，識字測驗分數與聽說讀寫測驗分數的相關係數一致高於完全計分，此顯示若是採部分計分，可提高題目鑑別度與測驗分數效度，且更能夠敏感反映 CSL／CFL 學習者的華語文能力。本研究結果顯示，理論上中文字層級是中文知識之核心成分，而在輔助華語教學實務方面，本研究發展的線上識字測驗可作為評估 CSL／CFL 學習者華語能力的有效工具。

關鍵詞：中文線上測驗系統、中文識字測驗、以中文為第二語言或外語學習者

1. 曾千芝，國立臺灣師範大學教育心理與輔導學系博士生
2. 張瓅勻，國立臺灣師範大學應用華語文學系助理教授
3. 張雨霖，國立臺灣師範大學教育心理與輔導學系助理教授
4. 陳學志，國立臺灣師範大學教育心理與輔導學系教授
通訊作者：陳學志；Email：chcjyh@gmail.com
地址：臺北市大安區和平東路一段 162 號國立臺灣師範大學教育心理與輔導學系
Literature Review

The blossoms of computer network technologies have changed the forms of Chinese language education and that made the distance education of Chinese developed fast and vigorously. However, a problem has arisen since people who learn Chinese as a second/foreign language (hereinafter referred to as CSL/CFL learners) come from all over the world and their Chinese competencies are quite different. Therefore, they may feel frustrated or miss the opportunity of progress if they are not assigned to a class at a suitable level for their abilities before classes start because of the difficult learning materials or the materials are too simple. In view of this, the demand of an online test for Chinese distance education has been arising.

The empirical research has discovered that character size is an important criterion for the growth of an individual’s four language abilities. Prior reading research has revealed that the size of the character knowledge (i.e. the number of Chinese characters that a person has acquired) is positively correlated to the level of reading comprehension (Ko, 2006; Lu, 2010; Tsai, 2010; Wu & Huang, 1998). Xu (2008) also discovered that the reading abilities of CSL/CFL learners usually depend on the number of Chinese characters they know. Moreover, with respect to listening comprehension, the higher the ability in character recognition, the better the performance in listening comprehension (Wu & Dai, 2007). In addition to correlations between character knowledge and language comprehensions (reading and listening), there are similar findings in the aspects of language production (speaking and writing). Prior research has shown that the level of word-recognition ability is positively correlated to the level of oral narrative ability (Nien, 2008) and written pragmatic skills (Tsai, 2010). It is also evident that the number of characters used in an article is an effective index for determining the writing ability of the writer (Chen, 2007).

In spite of previous empirical data shows that Chinese character recognition ability has a predictive power for Chinese performance, after reviewing the development of standard Chinese Character Recognition Tests carefully, we found that there is no Web-based Character Recognition Assessment that can be directly used as a ref-
ference for CSL/CFL placement tests. Currently, existing character recognition tests, targeted at native Chinese students in junior high schools and elementary schools, all adopt the traditional paper-based testing method which includes the Graded Chinese Character Recognition Test (Huang, 2001) and the Assessment of Chinese Character Lists (Hung, Wang, Chang, & Chen, 2008). By examining these two developed tests (please see the details are in the following section), we aimed to identify core Chinese character knowledge and scoring methods that are suitable for evaluating the Chinese competence of CSL/CFL learners.

1. Chinese character knowledge

The Graded Chinese Character Recognition Test and the Assessment of Chinese Character Lists are tackling different dimensions of Chinese character knowledge. Therefore, they adopt different recognition methods. In the Graded Chinese Character Recognition Test, students were asked to write down Zhuyinfuhao (group test) or name (individual test) the pronunciation of the given characters, emphasizing the connections between the forms and the pronunciations. For Chinese characters, one character corresponds to one syllable. Such a syllable may contain an onset, a rime, and a tone. For example, /m/ is the onset and /a/ is the rime for the Chinese syllable /ma/. Each syllable can be further differentiated by four tones: High flat, rising, low dipping, and falling. When the onset, the rime and the tone are all correctly recognized, this character is considered to be correctly answered. With respect to the Assessment of Chinese Character Lists, in addition to reporting the pronunciation of given characters, participants are asked to construct words with given characters, which emphasizes both the connections of form-pronunciation and form-meaning. However, constructing words with given characters requires CSL/CFL learners to owning a certain level of extra ability to write Zhuyinfuhao. Moreover, the test results may underestimate an individual’s character recognition ability, and thus decrease the validity of the test. Therefore, based on the task of reporting the pronunciation of given characters in the Graded Chinese Character Recognition Test, the present study has conducted a Web-based Character Recognition Assessment, and verified whether this web-based character recognition assessment could replace the traditional paper-based character recognition tests.
2. Scoring method

Both the Graded Chinese Character Recognition Test and the Assessment of Chinese Character Lists adopt a tone scoring method (i.e., a credit is only given when the tone, onset, and rime are correctly produced). However, this method may underestimate the learners’ real abilities. The reasons are as follows: There is a general consensus that perception proceeds production during L2 phonological learning (Escudero, 2006). Thus, if the learners cannot perceive sounds well, they cannot pronounce sounds correctly. Prior research has shown that lexical tones are particularly challenging for atonal language speakers to acquire (Kiriloff, 1969) because the suprasegmental characteristic of lexical tones is absent from the atonal languages. Moreover, learning to recognize tone requires learners adjusting their categorical perception which is formed by their first languages and which is a relatively stable speech perception. Since it is harder for atonal language speakers to perceive lexical tones, it would be more difficult for these learners to produce correct tones. Thus, non-tone scoring method for CSL/CFL learners may have the advantage of well discrimination that could make the online character recognition test much more sensitive towards the character recognition abilities of the participants. Therefore, in addition to the tone scoring method, this study also adopted a non-tone scoring method (ignoring the tone, the credit is given when the onset and the rime are correctly answered). Furthermore, the non-tone scoring method will be tested in the reliability and the validity analysis for checking if it is better than the tone scoring method or not.

In short, in order to conduct an efficient and effective online placement test, three aims are set up in this study: Firstly, to construct an online Chinese character recognition system that is suitable for CSL/CFL learners, including the functions that enabling the system to provide flexible choices to the examiners when they are compiling test papers, and to establish a friendly interface for CSL/CFL learners; secondly, to compile a Chinese character recognition test on this system to verify whether this Chinese character recognition assessment has good reliability and validity. Specifically, we would verify this assessment’s predictive effects on traditional listening, speaking, reading, and writing tests; thirdly, with respect to the scoring methods of this character recognition assessment, the non-tone scoring method would be verified.
in order to examine whether it is better than the tone scoring method in predicting the participants’ competency in Chinese listening, speaking, reading, and writing.

Method

The goal of this study is to develop an online placement test for CSL/CFL learners. In what follows, first, by performing item analysis, we conduct an online Chinese Character Recognition Test (website: http://140.122.109.78/Test/) with a set of 120 characters. Second, we perform the analysis for reliability and validity of the Chinese character recognition test.

1. Item analysis of the Chinese Character Recognition Assessment

In order to compile a test with 120 characters, this study first selected 160 high-frequency characters and then picked the test questions in accordance with the item analysis. The detailed description is as follows.

(1) Select items

This system has a built-in character database of 6,097 characters which is taken from the connex set consisting of 5,401 characters defined by the BIG-5 code and 5,656 characters concluded by the Chinese Knowledge and Information Processing Group of Academia Sinica (1993). In order to select items for the character recognition assessment, we referred to the grading method of the Graded Chinese Character Recognition Test (Huang, 2001). Based on the character frequency summed by the corpus-based frequency word count in statistics on the character frequency of journal vocabularies (Chinese Knowledge and Information Processing Group of Academia Sinica, 1993), the entire 6,097 characters in the character database were divided into 10 levels equally in this study. Level 1 contains the most frequent characters and level 10 contains the least frequent characters. After the procedure, 40 characters were randomly selected from each of the high-frequency levels one to four, which were the top 40.00% of characters with high character frequency in the list of 6,097 characters. That is to say, the items that were used in this study were randomly selected from 2,400 characters in the database and whose character frequency was ranked in the top
2,400. According to Huang (1994), the number of Chinese characters that adults need to know for daily lives is 2,328. The average character frequencies from level one to level four were 1362.93, 173.60, 61.76, and 25.72 times in every million characters, respectively. Consequently, a total of 160 characters were selected as the testing items for item analysis.

(2) Participants

The one hundred and six participants in the item analysis were foreign students who studied Chinese in Taiwan, with 50 males and 56 females. Their ages ranged from 18 to 25 with an average age of 20.08. Eighty-two of them came from America; twenty-one came from Europe and three came from Thailand.

(3) Research procedures

a. Read the test instruction

Before starting the test, the participants should read the instruction first: “The goal of this subtest is to estimate how many Chinese characters you know. Later, you will see a character on the screen, please use the mouse to click the keyboard and report the pronunciation of this character. In short, please tell me how to pronounce this character. Maybe you have not learned some of the characters, but try your best to answer every item to the best of your ability. If you want to alter the answer, please click “Eliminate the answer”, and please click “Send the answer” after you have confirmed the answer. You have 15 seconds to answer each item, and there are a total of 160 items.

b. Select the input method

The participants can choose to use Zhuyinfuhao or Chinese pinyin to describe the pronunciation of each item.

c. Five items for practice

Practice items include “木”, “七”, “人”, “不”, and “一”. None of them will appear in the following formal items.
d. Answer the formal items

Items are given in order of high-frequency characters to low-frequency characters in an easy-to-difficult order. Figures 1(a) and (b) show the answering interface, presented in both Chinese and English. When the items are finished, the system would give the information that the test has been finished. However, if 30 consecutive items are answered incorrectly, the test would be terminated earlier than it is supposed to be. It takes an average of eight minutes to finish the test.

![Figure 1(a) Answering Interface for Zhuyinfuhao](image1)

![Figure 1(b) Answering Interface for Chinese Pinyin](image2)

The scoring system for the test results includes both the tone scoring method and the non-tone scoring method, which both adopt dichotomous scoring. For either method, answering one item correctly would gain one point, and reporting a wrong answer would score zero. The total score is 160 points, and a higher score stands for higher character recognition abilities.

(4) Implementation of item analysis

Before the item analysis, multiple-reading characters (one such character) and phonograms that phonetic radicals share the same onset and rime with the exact character (three such characters) were excluded from the 160 characters. Multiple-reading characters refer to those characters with at least two different pronunciations. For example, character “給” (means give) can be pronounced as either gěi or ji. Because there are more than two accurate answers for the multiple-reading characters, it may
confuse the scoring and so these types of items were deleted. Finally, phonograms that phonetic radicals share the same onset and rime with the exact character were also excluded because Chinese phonograms contain a semantic radical and phonetic radical, which would give hints on the meaning and pronunciation of the character. For instance, for character “洋”, its phonetic radical “羊” has the pronunciation of yáng, and the character “洋” itself has the same pronunciation of yáng. Participants may guess the pronunciations of such characters during the test even if they actually do not recognize these characters. Thus, there were three such characters excluded from the 160 characters.

Afterward, the item analysis was implemented by the non-tone scoring method, and correlations between the credit of each item and the total credit were used as the basis for selecting items for the formal test paper. The top 30 characters in each level were the items for the formal test and the rest 36 items were deleted. The formal test paper contains 120 characters. Besides, the average number of strokes is 4.81, varying from 1 to 10 strokes and the average character frequency is an appearance of 435.4 times in every million characters.

Table 1 shows the results of the average item difficulty and the average item discrimination for the items under each character frequency level. The calculation of item difficulty is the passing ratio of this item. The results show that the higher the character frequency, the lower the difficulty level. In addition, the difficulties of the tone scoring method are higher than the non-tone scoring method. The calculation of item discrimination is PH-PL for this item, in which PH is the ratio of the correct answers of the group with a high score (top 25%) and PL is the ratio of the right answers of the group with a low score (bottom 25%). The results show that, regardless of each character frequency, the level of item discrimination of the non-tone scoring method is higher than that of the tone scoring method.
Table 1. Means of item difficulty and item discrimination under each character frequency

<table>
<thead>
<tr>
<th>Character frequency level</th>
<th>Characters for formal items</th>
<th>Non-tone scoring method</th>
<th>Tone scoring method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(n = 30)</td>
<td></td>
<td>.37</td>
<td>.26</td>
</tr>
<tr>
<td>2(n = 30)</td>
<td></td>
<td>.19</td>
<td>.12</td>
</tr>
<tr>
<td>3(n = 30)</td>
<td></td>
<td>.14</td>
<td>.09</td>
</tr>
<tr>
<td>4(n = 30)</td>
<td></td>
<td>.07</td>
<td>.05</td>
</tr>
</tbody>
</table>

2. Implementation of reliability

(1) One-week test-retest reliability

20 CSL learners were recruited with 6 (30.00%) males and 14 (70.00%) females, aged from 19 to 27 (mean age: 22.35 year). 17 (85.00%) participants came from Vietnam, 2 (10.00%) from Korea, and 1 (5.00%) from Japan. All of them have studied in Taiwan for three months. The participants were asked to rate current reading, writing, listening and speaking abilities for Chinese according to the scale: 1. Very poor; 2. Poor; 3. Fair; 4. Functional; 5. Good; 6. Very good; 7. Native-like. The result showed that the mean scores for reading, writing, listening and speaking abilities were 4.4, 3.9, 4.3, 4.0, respectively.

(2) Cronbach’s Alpha and split-half reliability

205 CSL learners were recruited with 81 (39.51%) males and 124 (60.49%) females, aged from 14 to 25 (mean age: 19.17 year). 104 (50.73%) participants came from Asia, 68 (33.17%) from North America, 26 (12.68%) from Europe, and 7 (3.41%) from other areas.

Regarding the participants’ language proficiency levels, this study adopts the reading and listening tests for criterion-related validity (for specific details, please refer to the section of criterion-related validity) to conduct an evaluation. As Table 2
shows, first of all, Chinese teaching experts were invited to classify each item into a specific CEFR level and then we categorized each participant into one of three groups with the different CEFR levels (below A1, A1, A2 and above A2) according to the participants’ pass rate of the item tests. In the reading section, 77 participants (37.56%) were grouped below the A1 of the CEFR, 55 were grouped (26.83%) in A1, and 73 (35.61%) were in the A2 and above A2 group. In the listening section, 36 (17.56%) were grouped below A1, 49 (23.90%) in A1, and 120 (58.54%) were grouped in A2 and above A2. Based on the questionnaires conducted by the Chinese learning institute, 161 participants (78.54%) learned Chinese before and the average time of learning Chinese was 3.13 years, 33 (16.10%) never learned Chinese before and 11 (5.37%) did not specify. Regarding the usage or access to Mandarin Chinese in their daily lives, 33 participants (16.10%) chose “Not at all”, 67 (32.68%) “Seldom”, 72 (35.12%) “Sometimes”, 19 (9.27%) “Very often” and the other 14 (6.83%) did not specify.

<table>
<thead>
<tr>
<th></th>
<th>Below A1</th>
<th>A1</th>
<th>A2 and above A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEFR level of reading</td>
<td>77(37.56%)</td>
<td>55(26.83%)</td>
<td>73(35.61%)</td>
</tr>
<tr>
<td>CEFR level of listening</td>
<td>36(17.56%)</td>
<td>49(23.90%)</td>
<td>120(58.54%)</td>
</tr>
</tbody>
</table>

3. Implementation of criterion-related validity: Graded Chinese Character Recognition Test

(1) Participants

Thirty CSL learners were recruited, 11 males and 19 females, aged from 18 to 23 (mean age: 19.7 year). 22 participants came from the US and eight came from other countries.

(2) Instruments

The Graded Chinese Character Recognition Test was compiled by Huang (2001). This scale aims to measure the students’ abilities of Chinese recognition from elementary schools and junior high schools. There are a total of 200 characters. The participants are asked to write down Zhuyinfuhao (group test) or name (individual
test) the pronunciation of each character. The total score is 200 points. The test-retest reliability coefficient for this scale is .81 to .95, the coefficient for internal consistency is .99, and the coefficient for split-half reliability is .99. The criterion-related validity for the stage of elementary school and the Reading Achievement Test for Elementary School Students (Chou, 1992) is .48 to .67, and the criterion-related validity for the stage of junior high school and Scholastic Aptitude Test (Lu, Lu, Fan, Chen, & Ou, 1993) is .49 to .64. With respect to the construct validity, the score of the Chinese character recognition assessment increases as the grade increases. Grades can explain 69.06% of the total variances of Chinese character recognition testing scores.

(3) Research procedures

Participants carried out the Web-based Chinese Character Recognition Test on computers first, and finished the test within an average time of eight minutes. One week later, they carried out the Graded Chinese Character Recognition Test in a group. Two pieces of A4 sized test paper were handed out to participants, and they were asked to write down the pronunciation of each character next to the character. They had 30 minutes to take the test.

4. Implementation of criterion-related validity: Chinese listening, speaking, reading, and writing tests

(1) Participants

Participants were the same as those in the reliability analysis.

(2) Instruments

The criterion for investigating the criterion-related validity of the Chinese Character Recognition Assessment are the Tests for Chinese listening, speaking, reading, and writing, which are the placement tests for the Chinese Class of the Extension School of Continuing Education of National Taiwan Normal University (2010). When compiling the test questions, two experts in Chinese pedagogy from the Department of Applied Chinese Language and Culture, National Taiwan Normal University were invited to examine the quality of the questions and whether the questions were consistent with the difficulty criterion of CEFR while compiling the test questions. The speaking, listening, and reading tests were computerized (website for tests:...)
http://140.122.109.78/Testonline/), and the writing test was implemented by a paper-based test. The total time needed for the whole test was about one hour.

The tests in Chinese listening, speaking, reading, and writing not only have expert validity, but have criterion-related validity for frequency of Mandarin experience (correlation coefficients range from .34 to .62, all \( p < .01 \)). With respect to reliability, the results show that the Cronbach’s Alpha for the Chinese reading test is .88, and .92 for the Chinese listening test, indicating good internal consistency. In addition, since the Chinese writing and speaking tests only contained two questions, and the difficulty level and the question type were greatly different from each other, so the analysis of the Cronbach’s Alpha value is not applicable.

The detailed information for each subtest is shown below:

a. Chinese Reading Test

The test content is confirmed in accordance with A1 and A2 levels of CEFR. There are a total of 36 questions. The question coverages are in A1, A2, and B1 levels which are taken up to 27.78%, 66.67%, and 5.56%, respectively. The questions are multiple-choice with four options. The test contains three categories: a semantic test, word/grammar test, and essay test, which take up 40.00%, 40.00%, and 20.00% of the total score of 100%, respectively. Group tests show that the average time for completing this subtest is seven minutes and thirty seconds.

b. Chinese Writing Test

This test has a total score of 100, and contains two parts: Comprehensive fill-in questions (40.00%) and an essay (60.00%). In the comprehensive fill-in part, five characters and five uncompleted sentences are given to students, and they need to place these five characters into the five sentences. In the essay part, students are asked to write an eighty-character themed essay with at least 5 words provided. The criterion of giving grade includes six equally weighted indices: Number of character, number of wrongly written character, usage of the words provided, grammar, vocabulary, and organization. The time for finishing this subtest is 30 minutes.

c. Chinese Listening Test

The test content is confirmed in accordance with the A1 and A2 levels of CEFR.
There are a total of 36 questions. The question coverages in A1 and A2 levels which are taken up to 27.78% and 72.22%, respectively. The questions are multiple-choice with four options. Participants need to select the most appropriate one in accordance with the context in the recording. The context can be divided into three categories: A listening test of single sentences, a listening test of conversations, and a listening test of an essay, comprising 37.50%, 37.50%, and 25.00%, respectively, of the total score of 100. The average time for finishing this subtest is eleven minutes and seven seconds.

d. Chinese Speaking Test

This test contains two parts: Theme speaking (50.00%) and telling a story based on the given pictures (50.00%). Taking “theme speaking” as an example, with the theme “What job do I want to do (我想做的職業工作?)”, participants need to make a statement targeted to this theme within one minute. The criterion of giving grade includes six equally weighted indices: Comprehension, logical inference, vocabulary, syntax, phonics, and sentence intonation. Two Chinese teachers mark each participant’s performance based on the five levels: A1, A2, B1, B2, and C1. Next, the scores given by two teachers are averaged and converted into a score in the 100-point system. The time for finishing this subtest is 10 minutes.

(3) Procedures

In the first stage, participants carried out listening, speaking, reading, and writing tests. First, they used computers for these four tests. After that, a paper-test for the writing portion was handed out, and the paper was collected after 30 minutes. The average time for finishing the whole test was about 60 minutes. In the second stage, participants completed the Web-based Chinese Character Recognition Test and finished the test within an average of eight minutes.

5. Implementation of criterion-related validity: Comparison between participants with different levels of CEFR (below A1, A1, A2 and above A2) on their character recognition test scores.

In order to prove that the character recognition test scores can tell between learners of different proficiency levels apart, this part adopts the response data (n =
205) from the reading test of the criterion-related validity tests. Chinese teaching experts were invited to classify each item into a CEFR level and then put each participant into one of three groups for different CEFR levels (below A1, A1, A2 and above A2) based on their performance to compare the differences in the character recognition tests of these three groups.

Results

We examined the reliability and validity of the Chinese Character Recognition test comprehensively by performing different types of reliability and validity analyses and using two scoring methods - tone scoring and non-tone scoring methods - under each analysis.

1. Results of reliability

One-week test-retest reliability coefficient was .84 while adopting the tone scoring method, and .83 while adopting the non-tone scoring method. As for the internal consistency reliability, the Cronbach’s Alpha for internal validity was .98 when adopting the tone scoring method, and .98 when adopting the non-tone scoring method. The study also verified the split-half reliability. The whole test was separated by odd number questions and even number questions to calculate the correlation coefficients between the two scores. The Pearson product-moment correlation coefficient for results gained using the tone scoring method is .97, and .97 for results gained by the non-tone scoring method. The effect sizes $r^2$ for the above results of reliability represented large effect sizes (Cohen, 1992). The above information indicates that the Web-based Chinese Character Recognition Assessment has good internal consistency.

2. Results of criterion-related validity

(1) Graded Chinese Character Recognition Test

When the tone scoring method was adopted, the correlation coefficient would be .97 ($p < .001$, effect size $r^2 = .94$, a large effect size) between the score for the Web-based Chinese Character Recognition Assessment and the score of the Graded Chinese Character Recognition Test. When the non-tone scoring method was adopted, the
correlation coefficient would be .91 ($p < .001$, effect size $r^2 = .83$, a large effect size) between the score for the Web-based Chinese Character Recognition Assessment and the score of the Graded Chinese Character Recognition Test. The score for the Web-based Chinese Character Recognition Assessment had a significantly positive correlation with that of the Graded Chinese Character Recognition Test.

(2) Chinese Listening, Speaking, Reading, and Writing Tests

Pearson correlation analysis was performed on the scores among the Chinese Character Recognition Test and the Chinese Listening, Speaking, Reading, and Writing Tests. Table 3 shows the mean scores and standard deviations for the accuracy. Table 4 presents the correlation matrix among the scores of the Web-based Chinese Character Recognition Test and the tests on Chinese Listening, Speaking, Reading, and Writing. We had two main findings. First, generally we found that there was a significantly positive correlation between the score for the Chinese Character Recognition Test and the scores for Chinese Listening, Speaking, reading, and Writing Tests (correlation coefficients range from .32 to .87, $r^2$ range from .10 to .76, effect sizes range from medium to large; all $p$s < .001) after examining all correlation coefficients. This finding suggests that people who score higher on the Chinese Character Recognition Tests also tend to score higher in listening, speaking, reading and writing tests.

Second, with regard to the correlations between scores of the Chinese Character Recognition Test and those of the listening, speaking, reading, and writing tests, we compared the results between using the tone scoring and non-tone scoring methods. The results reveal that for the reading score, the correlation coefficients between the scores gained by the tone scoring and the non-tone scoring methods were .73 and .78, respectively; for the writing score, they were .80 and .87, respectively; for the listening score, they were .49 and .56, respectively. Finally, for the speaking score, they were .32 and .35, respectively. These results suggest that compared to scores gained by the tone scoring method, the correlations between scores gained by the non-tone scoring method and the scores for the listening, speaking, reading, and writing tests were all higher although the correlation coefficients did not vary greatly.
Table 3. Mean scores and standard deviations of all test scores gained by participants

\( (n = 205) \)

<table>
<thead>
<tr>
<th>Test</th>
<th>Total score of the test</th>
<th>( M(SD) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score of Chinese Character Recognition Test</td>
<td>120</td>
<td>14.10(19.40)</td>
</tr>
<tr>
<td>(Tone scoring method)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score of Chinese Character Recognition Test</td>
<td>120</td>
<td>20.82(23.24)</td>
</tr>
<tr>
<td>(Non-tone scoring method)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score of Chinese Reading Test</td>
<td>100</td>
<td>51.03(20.90)</td>
</tr>
<tr>
<td>Score of Chinese Writing Test</td>
<td>100</td>
<td>34.64(28.80)</td>
</tr>
<tr>
<td>Score of Chinese Listening Test</td>
<td>100</td>
<td>61.11(23.92)</td>
</tr>
<tr>
<td>Score of Chinese Speaking Test</td>
<td>100</td>
<td>36.53(16.71)</td>
</tr>
</tbody>
</table>

Table 4. Correlation matrix scores for the Web-based Chinese Character Recognition Test and scores for the Chinese Listening, Speaking, Reading, and Writing Tests

\( (n = 205) \)

<table>
<thead>
<tr>
<th>Score of Chinese Character Recognition Test (Tone scoring method)</th>
<th>Score of Chinese Character Recognition Test (Non-tone scoring method)</th>
<th>Score of Chinese Reading Test</th>
<th>Score of Chinese Writing Test</th>
<th>Score of Chinese Listening Test</th>
<th>Score of Chinese Speaking Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.96**</td>
<td>.73**</td>
<td>.80**</td>
<td>.49**</td>
<td>.32**</td>
</tr>
<tr>
<td>1</td>
<td>.78**</td>
<td>.87**</td>
<td>.56**</td>
<td>.35**</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.82**</td>
<td>.60**</td>
<td>.44**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.60**</td>
<td>.37**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>.57**</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**\( p < .01 \)
(3) Comparison between participants with different levels of CEFR (below A1, A1, A2 and above A2) on their character recognition test scores.

Table 5 shows the sample size of participants in three CEFR groups and their means and standard deviations in the character recognition test (including tone scoring and non-tone scoring methods). The results show that the character recognition scores are positively associated with the group levels of CEFR no matter it is using tone scoring or non-tone scoring methods. Table 6 summarizes the results: Regardless of scoring methods, there are significant differences in character recognition scores among the participants in different CEFR groups (for tone scoring method: $F(2, 202) = 59.68, p < .001, \eta^2 = .24$, a medium effect size; for non-one scoring method, $F(2, 202) = 93.72, p < .001, \eta^2 = .27$, a large effect size). Furthermore, the Post Hoc tests revealed that the “A2 and above A2” group had higher character recognitions score than the “A1” group, and “A1” group had higher character recognitions score than the “below A1” group.

Table 5. Number of participants joining CEFR test of different levels (below A1, A1, A2 and Above A2) and Mean and SD of Character Recognition Test (including both tone scoring method and non-tone scoring method)

<table>
<thead>
<tr>
<th>CEFR levels</th>
<th>Tone scoring method</th>
<th></th>
<th>Non-tone scoring method</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Below A1</td>
<td>77</td>
<td>2.43</td>
<td>3.84</td>
<td>4.47</td>
</tr>
<tr>
<td>A1</td>
<td>55</td>
<td>10.11</td>
<td>12.50</td>
<td>16.25</td>
</tr>
<tr>
<td>A2 and Above A2</td>
<td>73</td>
<td>29.42</td>
<td>23.18</td>
<td>41.52</td>
</tr>
</tbody>
</table>
Table 6. ANOVA test results on the scores of character recognition test of participants joining CEFR test of different levels (below A1, A1, A2 and Above A2)

<table>
<thead>
<tr>
<th>Sources of variance</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean-square</th>
<th>F</th>
<th>p</th>
<th>η²</th>
<th>Post Hoc tests</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tone scoring method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>28508.81</td>
<td>2</td>
<td>14254.41</td>
<td>59.68</td>
<td>&lt;.001</td>
<td>.24</td>
<td>A2 and Above</td>
</tr>
<tr>
<td>Within Groups</td>
<td>48246.04</td>
<td>202</td>
<td>238.84</td>
<td></td>
<td></td>
<td></td>
<td>A2 &gt; A1 &gt; Below A1</td>
</tr>
<tr>
<td>Total</td>
<td>76754.85</td>
<td>204</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A2 and Above</td>
</tr>
<tr>
<td><strong>Non-tone scoring method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>53017.85</td>
<td>2</td>
<td>26508.93</td>
<td>93.72</td>
<td>&lt;.001</td>
<td>.27</td>
<td>A2 and Above</td>
</tr>
<tr>
<td>Within Groups</td>
<td>57137.82</td>
<td>202</td>
<td>282.86</td>
<td></td>
<td></td>
<td></td>
<td>A2 &gt; A1 &gt; Below A1</td>
</tr>
<tr>
<td>Total</td>
<td>110155.68</td>
<td>204</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A2 and Above</td>
</tr>
</tbody>
</table>

Note. The pairwise comparisons are significant at $p < .05$

**DISCUSSION**

The main goal of this study was to conduct an online character recognition test system for CSL/CFL learners. Some useful evidences have been gathered that contributes to three criterion-related validity. First, this study used the Graded Chinese Character Recognition Test (Huang, 2001) that also asks participants to report the pronunciation of a given character. The results suggest that there is a strong and positive correlation between the scores of the Web-based Chinese Character Recognition Test and those of the paper-based Graded Chinese Character Recognition Test. This supports that the Web-based Chinese Character Recognition Test is effective for assessing CSL/CFL learners’ character recognition abilities.

Second, the criterion used was the placement test for Chinese classes of the Extension School of Continuing Education of National Taiwan Normal University - Chinese listening, speaking, reading, and writing test. The results show that there was a significantly positive correlation (the correlation coefficients ranging from .32 to .87, all $ps < .001$) between the score gained in the Chinese Character Recognition Test and the score gained in the four language skills test. This result is consistent with previous studies; existing research indicates that the more the character knowledge, the better the performance in the listening, speaking, reading, and writing tests (Chen,
Thus, these results confirm the effectiveness of the scores using the Web-based Chinese Character Recognition Test as an indicator for assessing general Chinese competence.

Third, we compared the character recognition scores of the participants in the different CEFR groups. We found that the character recognition score can effectively distinguish the participants at the three different levels of CEFR (below A1, A1, A2 and above). This finding suggests that our online character recognition test can help to discriminate the learners of different proficiency levels.

The Chinese Character Recognition Assessment developed in this study contains two scoring methods: Tone scoring and non-tone scoring. Comparing these two scoring methods, we discovered that when the non-tone scoring method is adopted, the correlation coefficients between the score for the Chinese Character Recognition Test and the scores for the listening, speaking, reading, and writing tests are all higher than those coefficients when tone scoring method is adopted. The results indicate that the tone scoring method already possesses the function of discriminating the level of students’ Chinese competence given that the score for the character recognition test has a significantly positive correlation with the scores for listening, speaking, reading, and writing tests, and if the non-tone scoring method is adopted, the discrimination of items and the validity of testing scores will be enhanced.

What cause the different results of the two methods? We found that the difficulty level of the non-tone scoring method tends to be moderate (the passing ratio is close to .50). Therefore, it is equipped with relatively high item discrimination. Moreover, the two methods may reflect different levels of recognizing character performance: The non-tone scoring method requires learners recognizing correctly, while the non-tone scoring method requires learners not only recognizing correctly, but also producing correctly.

It is worth noting that Chinese is a tonal language, but we found the correlation coefficients between the score for the Chinese Character Recognition Test and the scores for the speaking tests is higher than the coefficients when tone scoring method is adopted. The cause probably would be the criterion of giving grade which includes comprehension, logical inference, vocabulary, syntax, phonics, and sentence intonation. Thus, the learners’ tonal ability should take little part in the scores for the spe-
On the other hand, we found that error ratio in the tone of CSL learners is significantly high. Every 3.1 items has one item in which the participants gave a correct answer for onset and rime but the wrong tone. The result highlights the difficulty of learning different tones. Empirical studies on Chinese teaching have shown that CSL learners with a different first language background are prone to tonal errors and that these errors are hard to correct (Chen, 2007; Gao, 2007; Gao & Lee, 2006; Liao & Liao, 2010). For instance, Liao and Liao (2010) found that French-speaking CSL learners are generally unable to master tone 2 and tone 3. The result highlights the need of strengthening the learning of tones in Chinese character teaching.

Three suggestions that will extend these results and their applications are proposed for future studies. First, the Chinese Character Recognition Test in this study was completed using either Chinese pinyin or Zhuyinfu. However, the phonetic system can be easily affected by the fact that Chinese phonograms can be pronounced in accordance with the pronunciation of their phonetic radicals. Although phonograms which phonetic radicals sharing the same pronunciation with the exact character have been excluded during the compilation of this test paper, as phonograms include the majority of Chinese characters, reaching about 80%, it is important not to ignore an assessment of 80% of the characters. Moreover, the exclusion could introduce the risk of under-estimation of the number of recognized Chinese characters while making partial judgment on Chinese character recognition. Therefore, it is expected that a Chinese pronunciation assessment based on automatic speech recognition will be developed in the future. Participants will be asked to pronounce the given character and construct a word, and the automatic speech recognition module will assess whether the pronunciation and the word constructed are correct or not. Second, the passing ratio for non-tone scoring method is 0.19, the passing ratio for tone scoring method is 0.13. These results demonstrated that the average item difficulty is quite low. Therefore, in the future, another assessment can be compiled using characters with comparatively high character frequency and low difficulty to evaluate the Chinese competency of different CSL/CFL learners. Third, the split-half reliability in this study is quite high, especially when the non-tone scoring meth-
od was adopted and the value reached .97. The high reliability suggests that, for future assessment, the length of the assessment can be shortened by 50%, and thus the time taken for testing can also be greatly reduced, enhancing the economic benefits of time.

Acknowledgements

This research is partially supported by the “Aim for the Top University Project” and “Center of Learning Technology for Chinese” of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, R.O.C. and the “International Research-Intensive Center of Excellence Program” of NTNU and Ministry of Science and Technology, Taiwan, R.O.C. under Grant no. MOST 105-2911-I-003-301.
References


