Behavior Research Methods (2025) 57:155
https://doi.org/10.3758/s13428-025-02680-9

ORIGINAL MANUSCRIPT

Chinese character size test: Test development, validation, and standards-referenced norms for Chinese primary students

Yan Li[1] · Yi Wei[2] · Hong Li[3]

Accepted: 31 March 2025 / Published online: 25 April 2025

Abstract

Recognizing Chinese characters is a fundamental skill that is crucial for later reading development and basic educational achievement in China. However, there is currently a lack of published tests or norms that provide adequate feedback on children's character size development. To address this gap, we developed the Chinese character size test (CCST) for Mandarin Chinese students in grades 1–6. To address the need for resource-efficient assessment, longitudinal tracking precision, and curriculum-aligned interpretation, we applied rigorous test development processes, vertical test equating, and the item response theory framework to measure children's character size development throughout primary grades. A comprehensive evaluation of the CCST's psychometric properties yielded satisfying results, including item statistics, test reliability, inferential score reliability, criterion-related validity, and empirical validity. Normative data from a representative sample of 7,459 primary school students in Beijing were analyzed to construct the norm-referenced and criterion-referenced character size scores. The results indicated that the mean character sizes of primary students in grades 1–6 were 1,227, 1,898, 2,422, 2,722, 2,932, and 3,060 characters, respectively, and that approximately 5.1% of the students in grades 3–6 failed to achieve the required character size level by the national curriculum criterion. In conclusion, the CCST is a child-friendly, highly interpretable, and open-access instrument with strong psychometric quality and comprehensive scoring feedback. This work would interest a wide range of users, including researchers, educators, and practitioners.

Keywords Chinese character size · Item response theory · Norms · Test development and validation · Chinese character recognition · Primary students

Introduction

Recognizing printed characters is a fundamental skill, and children who fail in this skill are more likely to struggle with reading, have learning difficulties, and drop out—all at a high cost to society. Therefore, we look to the early stages of basic education to ensure that native Chinese children develop proficient character recognition skills (McBride-Chang & Chen, 2003). The Chinese Language Curriculum Criterion for Compulsory Education (2022 Edition) (hereafter referred to as the curriculum criterion) sets quantifiable benchmarks for the teaching and learning of Chinese characters. Namely, students who complete the second, fourth, and sixth grades of primary school should be able to recognize 1,600, 2,500, and 3,000 commonly used Chinese characters, respectively (Ministry of Education, 2022). This quantifiable approach is grounded in two key considerations: (1) the pragmatic nature of Chinese characters as morphemes, that is, basic meaning-bearing units that independently or combinatorially form words (Shu et al., 2003); and (2) empirical evidence demonstrating that mastery of 3,000 characters enables comprehension of approximately 99% of Chinese texts (Li et al., 2022a, 2022b), establishing it as the threshold for independent reading as mandated by the national curriculum (Ministry of Education, 2022). Therefore, a valid, interpretable, and accessible tool for measuring Chinese

✉ Hong Li
psy.lihong@bnu.edu.cn

[1] School of Teacher Development, Shaanxi Institute of Teacher Development, Shaanxi Normal University, Xi'an, China

[2] China National Children's Center, Beijing, China

[3] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Institute of Children's Reading and Learning, Faculty of Psychology, Beijing Normal University, Beijing, China

character size is highly desirable for research and teaching purposes.

Chinese character recognition and its assessments

How do children develop and reach the milestones of Chinese character recognition (CR)? CR is widely defined as decoding the pronunciation (phonology), shape (orthography), meaning (semantics), and use of a Chinese character in different situations (Pan et al., 2021; Perfetti et al., 2005; Shen, 2004; Tseng et al., 2016; Yeh et al., 2017). The Chinese language is a logographic writing system, and there is often an opaque correspondence between phonology and orthography (Chang et al., 2016). Therefore, although preschoolers can already understand and speak some characters in their daily lives, only the relationships between pronunciation and meaning are accumulated (Hao et al., 2008). It is in primary school that native Chinese children explicitly learn to recognize written characters in Chinese language courses and build up a large number of unified mental lexicons that link the form, sound, meanings, and uses of characters (Duff & Brydon, 2020; Li et al., 2022a, 2022b). CR shows rapid growth in children's primary school years (Cai et al., 2022). It has been regarded as one of the best early predictors for identifying children who later manifest the diagnosis of dyslexia (Guan et al., 2020; Ho et al., 2004; McBride-Chang et al., 2011), and a strong predictor of later educational outcomes (Bleses et al., 2016; Zhang et al., 2018). Therefore, CR is increasingly being measured in psycholinguistic research and educational practice, especially after some review and meta-analytical studies suggested that direct character instruction could significantly close the reading achievement gap (Catts, 2018; Hattie & Yates, 2013).

However, few large-scale instruments have been developed to assess CR in Chinese children, and their scoring interpretations of character size are even rarer. Past efforts to measure CR can be classified into three types: phonological, semantic, and combined test formats. Phonological test formats usually require participants to name a list of characters aloud to a professional subject (Leung et al., 2008; H. Li et al., 2012; Liu et al., 2007; Pan et al., 2011; Perfetti et al., 2005; Shu et al., 2008; Wei et al., 2014) or to write the pronunciations of the target characters in Pinyin or Zhuyin Fuhao (Liao et al., 2008; Tseng et al., 2016). In addition, the semantic format assesses participants'understanding of character meaning by verbally forming words or phrases using the target characters, as exemplified by the Elementary School Student Character Size Test (Wang & Tao, 1996). This standardized test assembled characters into parallel booklets stratified by frequency and grade level, covering all 3,500 characters from the *List of Commonly used Chinese Characters in Modern Chinese* (1988 edition), providing direct feedback on character size attainment. However, its exhaustive coverage and paper-and-pencil format imposed substantial response burden on young learners. The combination format employs various item types, such as writing down the characters, their pronunciation, homophones, and forming words (Ho et al., 2007; Hung et al., 2008; Wen et al., 2015). These tests generally showed high reliability and supported the unidimensionality hypothesis of Chinese character recognition. Zhang and colleagues (2022) validated the interchangeability of three CR test formats (phonological, semantic, and phonological-semantic) among Chinese-as-a-second-language learners. Their findings revealed that semantic tasks demonstrated superior psychometric properties for learners from alphabetic L1 backgrounds, while phonological-semantic formats showed stronger validity for logographic L1 learners. This underscores the importance of tailoring CR assessment formats to the linguistic profiles of learners. However, these CR tests mainly took the form of one-to-one verbal tests or paper-and-pencil tests, which are very demanding on the subjects and scoring. Therefore, it is worth exploring a more child-friendly and convenient way of testing.

Given the consensus on learning-oriented assessment to improve learning outcomes (Zeng et al., 2018), there is growing interest in more accurate and informative ways to interpret scores. Formative assessment of children's Chinese character size is essential and ongoing throughout their primary school years (Cai et al., 2022). However, conventional feedback of CR tests often employs raw scores (i.e., the number of correct answers and percentiles) generated by a limited number of items and participants under the classical test theory framework (Ho et al., 2007; Hung et al., 2008; Li et al., 2012; Liu et al., 2007; Pan et al., 2011; Perfetti et al., 2005; Shu et al., 2008; Wang & Tao, 1996; Wei et al., 2014; Wen et al., 2015). Individual statistics of such scoring are highly dependent on the sampling of test takers and the assembly of items, resulting in incompatible scores for growth measurement and limited classroom application. Thus, the methodological challenges of measuring Chinese character size lie in providing a more comprehensive and nuanced understanding of the innate nature of character recognition and acquisition, and how to provide curriculum-integrated and interpretable scorings for teachers to identify students who are significantly below benchmarks.

Item response theory framework

Item response theory (IRT) refers to a family of modern psychometric models for the probabilities of varying item responses as a function of item and person characteristics (Embretson & Reise, 2000). IRT item statistics are independent of the groups from which they are estimated, allowing scores to be standardized across different measurement

scales. Because of its strong advantages in reliability estimation, measurement invariance, and test equating, IRT is a widely accepted framework in psychological and educational measurement (Schneider et al., 2022). For example, Beglar (2010) adopted the Rasch model to validate a 140-item form of the Vocabulary Size Test, and these IRT-calibrated items could be flexibly combined for various measurement demands and participants, with an accurate and comparable scoring of written receptive knowledge that related to 14,000 English words. Another study showed that an IRT-like normal ogive formula could explain the highest variance proportion of the simulated Chinese character size, outperforming those calculated by the traditional corpus-based proportional inference methods (Wen et al., 2020). Therefore, the IRT application is needed in this study for optimal measurement accuracy and a better understanding of children's CR development.

However, the reliable and valid estimation of latent abilities tells only one side of the story, and the estimation of their inferential scores is the other. Since students' character size is often attractive to educators and parents, the present work responds to their needs by developing a criterion-referenced and norm-referenced test that is representative of both teaching materials (character pools) and learners (primary school students). We carefully followed the Chinese curriculum criterion and the current textbooks on how to design, sample, and assemble what should be tested for a student at each grade level. Traditionally, based on the commonly used Chinese character list in the curriculum criterion, Chinese language textbooks list the characters to be taught and studied in each school semester. The current Chinese language textbooks published by the People's Education Press (hereafter referred to as the unified textbooks because they have been popularized exclusively in mainland China since 2019) also provide new character lists, which have improved the arrangement and quantity of characters in each school semester (Li et al., 2022a, 2022b). Therefore, we proportionally sampled the characters from these lists to ensure the content and ecological validity of the test, and to support teaching decisions. In addition, the character size norm for real native Chinese children has not yet been established, making it difficult to increase stakeholder confidence for large-scale applications. To our knowledge, this study provides the first Chinese character size norm for Mandarin-speaking students at different stages of development.

Considering that approximately 20% of the world's population are Chinese speakers (Su et al., 2022), and the need for character instruction is particularly felt at the primary level, the present study aimed to develop an innovative tool, the Chinese character size test (CCST), for Mandarin Chinese primary school students. Specifically, this study aimed to apply the item response theory framework and test characteristic functions to (1) develop and construct the item bank of the CCST according to the curriculum criterion, the unified textbooks, and Messick's validity framework (Messick, 1990); (2) provide sufficient validation evidence of test reliability and validity, and evidence of the inferential accuracy of the Chinese character size scores; and (3) construct both norm-referenced and criterion-referenced norms using a representative large-scale sample of 7,459 students in grades 1–6. In addition, because of the rapid growth of character size throughout the primary school years, a multilevel booklet design and vertical equating are used to balance item content and difficulty across grade levels, thereby improving test accuracy and item bank efficiency. These processes can provide novel and valid assessment tools and new insights into children's character size development and instructional needs.

Methods

Test development

Following Schmitt and colleagues (2020), the CCST was developed through rigorous and systematic processes, including item format design, Chinese character pool and sampling, item development, and pilots.

Item format design As discussed, the item format should be child-friendly to understand and respond to, and convenient for large-scale investigation. Therefore, a novel multiple-choice test format is proposed (see Fig. 1), in which students listen to audio pronunciations of a Chinese character and its forming word, and distinguish the correct character from three distractors. The auditory word stems (e.g., "月亮 yue4 liang") are used to approximate character usage in a real-life language context. To control for the impact of word complexity on character recognition performance, auditory word stems were selected from high-frequency words that include the target character, avoiding homophones to reduce vocabulary interference. The distractor characters are morphologically related, semantically related, and homophonic to the given character. This allows the correct response to represent the student's full recognition of the targeted character since the student must have formed the phonological, morphological, and meaningful association to exclude all distractors. To reduce the phenomenon of forced guessing in multiple-choice items (Stoeckel et al., 2016), a fixed "I don't know" option is also presented in each item.

The Chinese character pool and sampling The total pool consists of 3,500 common characters based on the national Chinese curriculum criterion (Ministry of

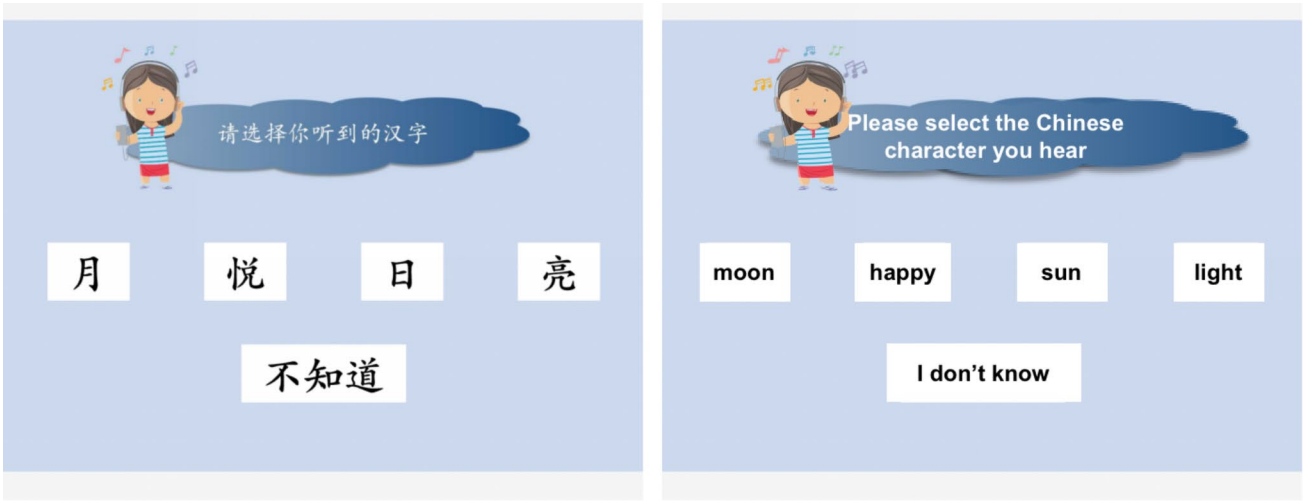

Fig. 1 A sample item from the CCST program. The stem is an automatically played audio, such as 月, 月亮的月[yue4, yue4 liang de yue4] (moon, moon in moon light). The options are 月 [yue4] (moon, correct), 悦 [yue4] (happy, homophonic), 日 [ri2] (sun, morphologically related), and 亮 [liang4] (light, semantically related). A fixed 不知道 (I don't know) option is also provided

Education, 2022). According to the unified textbooks, these 3,500 characters can be divided into 11 levels (i.e., the school semester of learning the character). By controlling the character level and character frequency in the CCL Corpus of Peking University (Zhan et al., 2019), 820 characters were sampled as target items using the stratified matrix sampling method.

Item development Test items were developed by a psychometrician and two graduates majoring in psychology. Characters that cannot form words were excluded. Two Chinese language teachers were invited for content validation. They revised item wordings and excluded inappropriate and rarely used items, such as 一[yi1] (one) and 埠[bu4] (cities that trade with foreign countries). After these procedures, 550 items were compiled, all of which had good face validity and consistency with the national curriculum.

Pilot test A pilot test was administered to 1,628 students in grades 1–6 in a primary school in Shaanxi Province, China. Items were assembled into seven linear test booklets for students in grades 1–6 and grade 3 (as enhanced anchor), using a preliminary design of vertical equating. Each booklet contained 150 items, with 50–100 items in common between adjacent booklets. Cronbach's alpha reliability for the seven booklets ranged from 0.88 to 0.95, with an average of 0.93. After piloting, 525 items with acceptable discrimination (item–total correlation > 0.20) were retained in the item bank.

Test administration

Norm sample

The norm sample consisted of 7,459 primary school students from 20 schools using the probability proportional to size (PPS) sampling method, representing approximately 980,000 primary school students in Beijing as a whole. That is, based on the geographical location and the number of students, 20 sample schools (nine central urban primary schools, six suburban primary schools, and five rural schools) were selected proportionally from four districts of Beijing. Then, students from two classes in grades 1–6 were randomly selected from the sample schools to participate in the assessment. The formal norm sample included 3,365 urban students, 2,343 suburban students, and 1,751 rural students. The number of students in each grade ranged from 1,200 to 1,279, with 51.4% being boys. (See Supplementary Table 1 for the detailed norm sample distribution.)

Table 1 The representativeness and the quality of the final item bank

Character levels	Number of character pool	Number of items	Sample ratio	Mean item difficulty	Mean item discrimination
1	300	23	7.67%	− 2.95	1.45
2	400	35	8.75%	− 1.38	2.34
3	450	57	12.67%	− 1.22	2.11
4	450	56	12.44%	− 1.08	2.12
5	250	37	14.80%	− 0.96	2.06
6	250	56	22.40%	− 0.82	1.96
7	250	40	16.00%	− 0.61	1.85
8	250	38	15.20%	− 0.24	1.89
9	200	36	18.00%	− 0.59	1.83
10	200	25	12.50%	− 0.5	2.09
11	500	122	24.40%	0.13	1.66
Overall	3,500	525	15.00%	− 0.72	1.92

Character levels 1–10 refer to the first to 10 th semesters of learning the common Chinese characters according to the unified textbook, and level 11 is assigned to the remaining common characters that are not directly taught at the primary school level. Several repeatedly taught polyphonic characters are classified according to their first appearance

Test instruments

The Chinese character size test (CCST) is an open-access standardized test that measures children's character recognition, the character size of simplified Chinese. The formal CCST includes six booklets respectively designed for primary school students in grades 1–6. Each booklet consists of 150 multiple-choice and dichotomously scored items that require students to select the correct Chinese character from homophonic, morphologically related, or semantically related distractors based on audio materials.

The CCST booklets were assembled using the vertical equation design. Each booklet includes 60 to 75 item characters learned during the current grade, 0 to 15 characters from previous grades, and 60 to 90 characters from the next grades. There are at least 40 anchor items between adjacent booklets, allowing for comparable test scores across test forms and grades. Moreover, this test assembly design can reduce the probability of students answering items that are too difficult or too easy, and improve the overall test discrimination. The CCST has no time limit. The average test time was 11.13 min, and 98.61% of students completed the test within 16 min. Therefore, the CCST can be easily administered to primary school students in a group setting.

The Diagnostic Chinese Reading Comprehension Assessment (DCRCA) is a validated reading comprehension assessment for Chinese primary school students in grades 2–6 (Li et al., 2021; Li et al., 2023). Items are questions on students' comprehension of short literacy or practical texts, in a dichotomously scored and multiple-choice format. The DCRCA contains three separate booklets for students in grades 2, 3–4, and 5–6, respectively. Each booklet contains 16 items to diagnose students'mastery of six reading attributes. Cronbach's alpha for the DCRCA ranged from 0.72 to 0.82. The average diagnostic reliability of the DCRCA was also above acceptable levels, namely 0.83 at the attribute level and 0.61 at the pattern level.

Test administration

The study was conducted during a large-scale reading education program in Beijing. The tests were conducted via an online website in the school computer classrooms under the supervision of subjects and Chinese/computer teachers. Before the formal tests, the test program displayed a standardized test instruction and two practice items to ensure that all participants understood the test requirements. Participants in grades 1–6 completed the formal CCST, and immediately thereafter, participants in grades 2–6 completed the DCRCA. Finally, we collected the most recent midterm exam scores of 243 students in grades 2–6 on Chinese language courses as the test criterion. One month after the tests, all sample schools received an analysis report about students'reading development and corresponding advice on teaching and reading strategies.

Data analysis

Data analysis included test quality validation, character size score estimation, and norm construction. First, the test quality validation was based on a two-parameter item response theory (IRT) model using the "mirt" package in R (Chalmers, 2012). Second, character size scores were estimated using the test characteristic function of IRT, which considers the character size as a logistic function of the Chinese character recognition ability and the item bank parameters (Tseng, 2013), so that

$$character\ size_i = \frac{\sum_{j=0}^{J} \mathrm{P_j}(\theta_\mathrm{i})}{J} \times \mathrm{T},$$

where $P_j(\theta_i)$ represents the item characteristic function of the two-parameter IRT model, which is the expected probability of the examinee i with the latent CR ability θ_i to respond correctly to the jth item. J is 525, which represents the number of items in the item bank. As alluded to above, the 525 items can be mapped into the 3,500 common characters listed in the curriculum criterion, so T is set to 3,500, representing the number of Chinese characters in the item pool. Thus, by summing the expected probabilities of correct responses $P_j(\theta_i)$ for all 525 items, the expected number of correct recognizing the entire item pool (3,500 characters) can be inferred proportionally. Essentially, the reliability of inferred scores can be addressed by determining confidence intervals for estimates of latent ability. Therefore, the 95% confidence intervals (CI) of character $\mathrm{size_i}$ are approximately transformed from the confidence interval of latent ability θ_i, using its standard error of measurement (SEM).

Finally, routine and essential uses of the Chinese character size score are to show how a score compares with those of other students and compares with cutoff point scores that distinguish character levels according to the curriculum criterion. Therefore, this study constructed standardized norms for primary school students in Beijing. Grade-level norms, gender norms, and urban–rural norms (urban districts, suburban districts, and rural areas) are introduced.

Results

Quality analysis of the item bank

The representativeness and the quality of the final item bank were examined. As shown in Table 1, the final item bank of 525 items sampled 15.0% of the total character pool, with the proportion of each character level ranging from 7.67% to 24.4%. Apart from the reasonable sampling ratio, the item bank and the overall pool of 3,500 characters showed similar means and standard deviations in important character features such as character frequency, contextual diversity, number of meanings, number of strokes, number of parts, and number of pronunciations (see Supplementary Table 2 for details), suggesting a good representativeness of the item bank.

The item bank was then vertically calibrated and analyzed using a two-parameter logistic IRT model, as it provided better fit indices than the Rasch or three-parameter logistic IRT model. As shown in Table 1, the mean item difficulties increased with character level, while item discrimination remained stable across levels, indicating that the vertical calibration of the items was reliable. The mean item difficulty of the 525 items was −0.72, ranging from −5.26 to 3.81, thus covering the range of CR skills of primary school children [− 4.37, 3.16]. The mean item discrimination parameter was 1.92, ranging from 0.12 to 4.43, indicating satisfactory item discrimination. Therefore, we ensured that the item bank of the CCST was in line with the Chinese curriculum in mainland China and could indicate the real status of learning and teaching Chinese characters at each grade level.

Test reliability and validity

The Cronbach's alpha reliability of the five formal CCST booklets was 0.93, ranging from 0.91 to 0.95. Figure 2A shows how much information the five booklets provided at various locations along the latent continuum of CR. The test information curves were above 9.70 for the students whose CR values were in the range of [− 3.51, 2.16], and their peaks gradually moved toward higher-level students, indicating that the CCST provides grade-appropriate precision for students at all elementary grade levels. Figure 2B shows the test information curve of the total item bank. According to Embretson and Reise (2000), the reliability of the full item bank is greater than 0.90.

Essentially, the reliability of the inferred scores can be addressed by their confidence intervals. Figure 3 plots the character size and its 95% CI as a function of CR, showing that if the same student were to take the assessment again, there is a 95% probability that the character size score would fall within a small interval of 162.50 adjacent characters on average. This suggests that the test scaling and inferred scores of the CCST are also highly reliable.

Criterion-related validity was tested by examining the correlations between the character size scores and children's reading ability and Chinese language achievement. As shown in Table 2, the CCST was positively correlated with reading comprehension, $r = 0.42$–0.64, $p < 0.001$, and Chinese language, $r = 0.43$–0.71, $p < 0.01$.

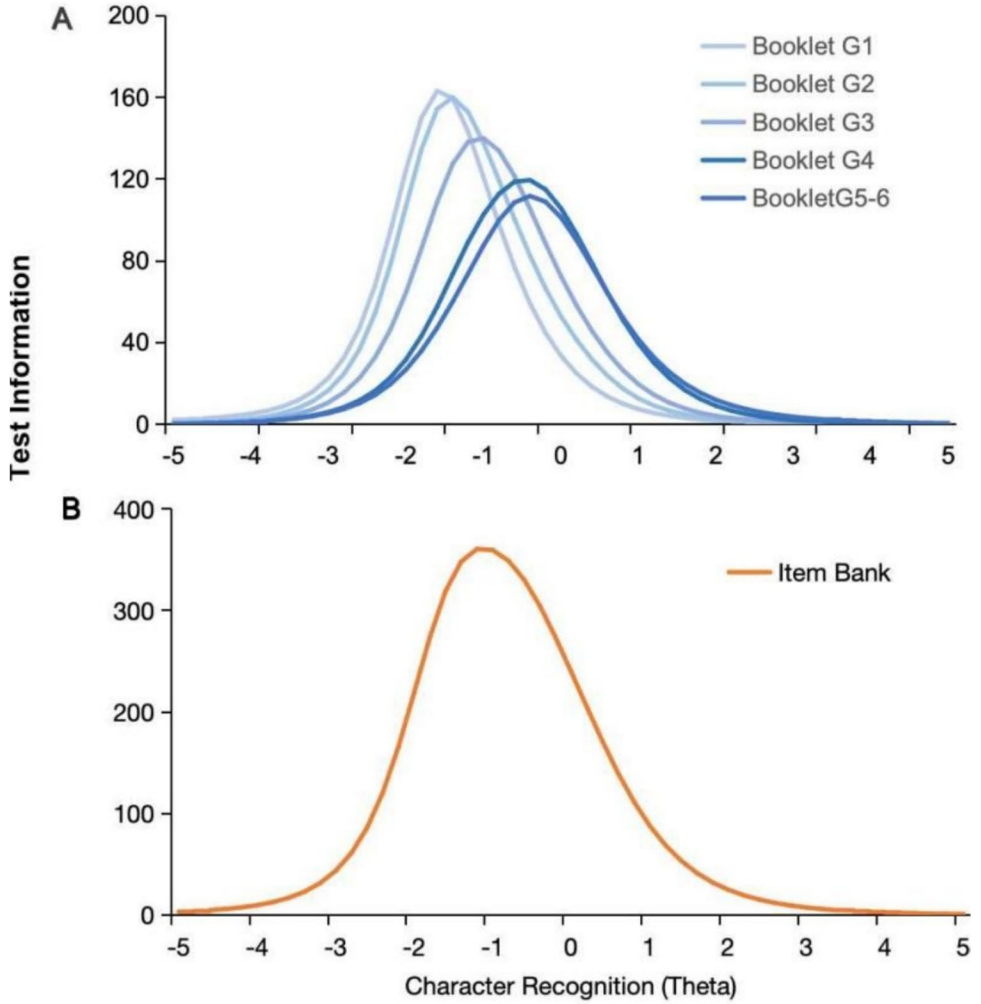

Fig. 2 Test reliability of the character size assessment. **A** Information provided by the five booklets. **B** Total item bank

This indicates that the CCST has good criterion-related validity and is a good predictor of reading ability and language scores.

Development of norms

Another important task of this study was to construct the standard-referenced norms for primary school students in China. Therefore, statistical scores including means (*M*), standard deviations (*SD*), and confidence intervals (CI) were reported for the total norm sample of 7,459 students and for students with different grade levels, gender, and regions. As shown in Table 3, the mean character sizes in grades 1 to 6 are 1,226.8, 1,898.5, 2,421.8, 2,721.9, 2,931.7, and 3,059.8 characters, respectively, with the growth trend from fast to slow. In addition, the mean character sizes in grades 2, 4, and 6 were all above their curriculum benchmarks of 1,600, 2,500, and 3,000 characters.

A three-way analysis of variance (ANOVA) of grade × region × gender showed significant differences among students in different grades, $F(5, 7{,}423) = 2{,}595.32$, $p < 0.001$, $\eta^2 = 0.64$; students in different districts, $F(2, 7{,}423) = 39.60$, $p < 0.001$, $\eta^2 = 0.01$; and between boys and girls, $F(1, 7{,}423) = 34.33$, $p < 0.001$, $\eta^2 = 0.005$. In addition, there was a significant interaction between district and grade level, $F(10, 7{,}423) = 14.50$, $p < 0.001$, $\eta^2 = 0.02$, while the remaining interactions were not significant. Further multiple comparisons showed that there were significant differences between all grades 1 to 6 ($p < 0.001$). Students in central urban districts had significantly better character size than students in suburban counties and rural areas ($p < 0.001$), and students in suburban counties had significantly better character size than rural students ($p < 0.01$). This indicates that the CCST

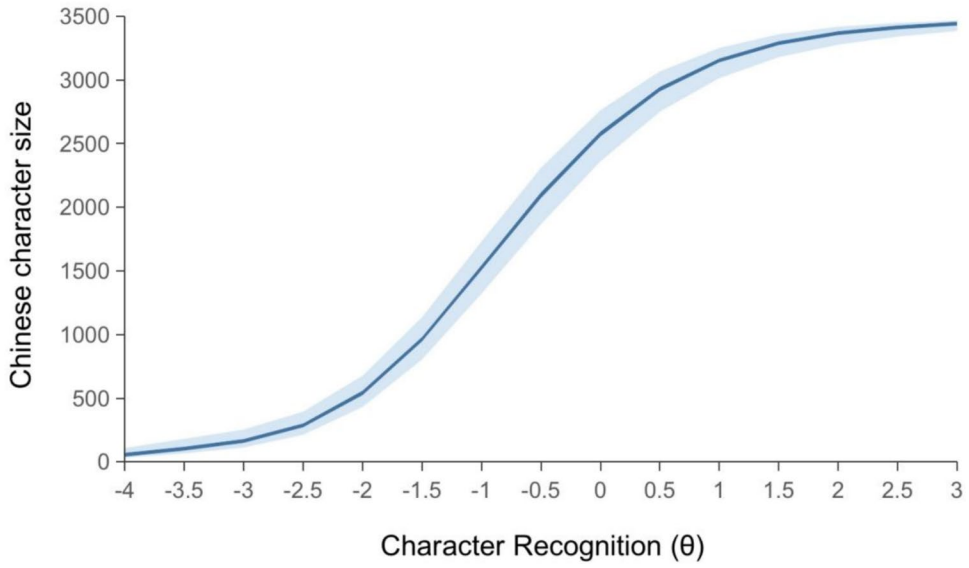

Fig. 3 The 95% CI of the Chinese character size scores

Table 2 Criterion-related validity of the CCST

Criterion variable	G2	G3	G4	G5	G6
Reading ability	0.64***	0.48***	0.55***	0.54***	0.42***
Chinese achievement	0.43**	0.71**	0.49**	0.53**	0.71**

Reading ability and the Chinese language achievement were measured by the DCRCA and Chinese mid-term exam scores, respectively. ** $p < 0.01$, *** $p < 0.001$

can effectively differentiate between grade, region, and gender factors, and has good empirical validity.

From a criterion-referenced perspective, students are grouped into four levels based on the national curriculum criterion. Level 1 represents a character size of [0, 1,600), level 2 represents [1,600, 2,500), level 3 represents [2,500, 3,000), and level 4 represents [3,000, 3,500). Figure 4 presents the distribution of students'character size levels in each grade of the norm sample, showing that the majority distribution

Table 3 The total, gender, and urban–rural norms of character size in each grade

Norm type	Stat	G1	G2	G3	G4	G5	G6	Average
Total	*M* (*SD*)	1,226.8 (768.9)	1,898.5 (599.6)	2,421.8 (492.6)	2,721.9 (375.0)	2,930.7 (307.8)	3,095.8 (244.2)	2,390.4 (812.5)
	95% CI	[1,183.6, 1,270.0]	[1,865.1, 1,931.9]	[2,394.4, 2,449.1]	[2,700.6, 2,743.1]	[2,913.8, 2,947.6]	[3,082.4, 3,109.2]	[2,372.0, 2,408.8]
Boys	*M* (*SD*)	1,223.6 (767.7)	1,857.4 (610.5)	2,382.3 (495.1)	2,693 (404.3)	2,902.1 (319.6)	3,045.1 (268.6)	2,353.6 (810.2)
	95% CI	[1,163.5, 1,283.7]	[1,809.7, 1,905.0]	[2,344.9, 2,419.6]	[2,660.5, 2,725.4]	[2,877.5, 2,926.7]	[3,024.3, 3,065.8]	[2,327.9, 2,379.2]
Girls	*M* (*SD*)	1,230.2 (770.9)	1,941.4 (585.4)	2,468.7 (485.9)	2,750.7 (341.3)	2,960.2 (292.4)	3,147.9 (203.9)	2,429.4 (813.2)
	95% CI	[1,167.7, 1,292.6]	[1,894.7, 1,988]	[2,428.7, 2,508.6]	[2,723.3, 2,778.0]	[2,937.3, 2,983.1]	[3,131.9, 3,163.8]	[2,402.9, 2,455.9]
Central urban	*M* (*SD*)	1,143.1 (661.9)	1,916.3 (565.1)	2,544.9 (452.3)	2,849.7 (310.3)	3,006.6 (275.7)	3,128.9 (216.3)	2,428.7 (827.8)
	95% CI	[1,088.3, 1,197.9]	[1,869.4, 1,963.2]	[2,507.7, 2,582.0]	[2,823.5, 2,875.9]	[2,984.4, 3,028.8]	[3,110.5, 3,147.3]	[2,400.7, 2,456.7]
Suburban	*M* (*SD*)	1,361.5 (872.8)	1,907.8 (612.5)	2,345.9 (477.0)	2,624.3 (368.3)	2,901.8 (287.6)	3,076.5 (254.8)	2,379.4 (784.4)
	95% CI	[1,272.8, 1,450.1]	[1,847.9, 1,967.8]	[2,297.9, 2,393.9]	[2,587.2, 2,661.5]	[2,873.2, 2,930.4]	[3,051.8, 3,101.2]	[2,347.6, 2,411.2]
Rural	*M* (*SD*)	1,214.6 (798)	1,848.6 (646.5)	2,280.4 (530.4)	2,606.8 (420.2)	2,815 (350.8)	3,066.4 (266.7)	2,331.5 (816.1)
	95% CI	[1,120.7, 1,308.5]	[1,772, 1,925.2]	[2,219.4, 2,341.3]	[2,557.3, 2,656.4]	[2,774.6, 2,855.3]	[3,037.5, 3,095.2]	[2,293.2, 2,369.7]

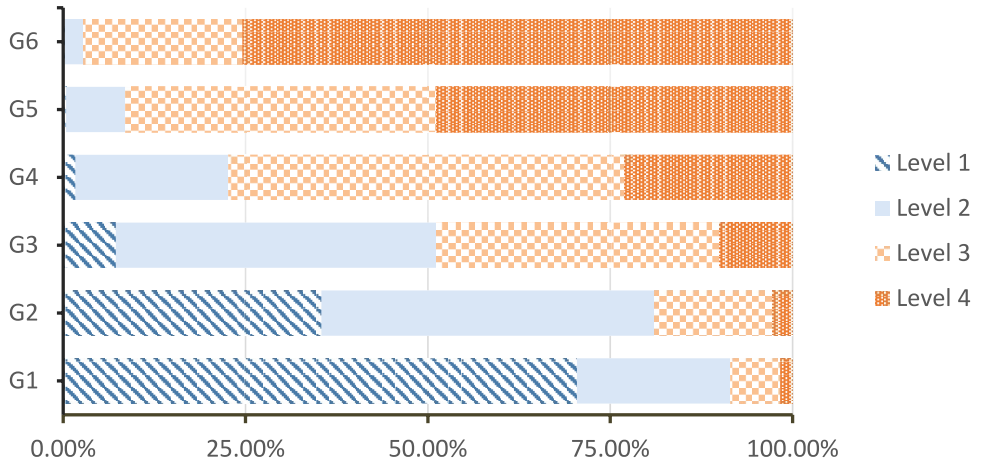

Fig. 4 The proportion of character size levels in each grade. *Note:* Students in grades 1–6 are grouped into four levels based on the curriculum benchmarks

of students'character levels is largely consistent with the curriculum benchmarks. For example, among the first-grade students, 70.44% of them were at level 1, and only 21.10% of them reached level 2. However, among the second- and third-grade students, level 2 became the most dominant type, with its percentage rising to 45.65% and 43.95%, respectively, which is in line with the curriculum benchmark that students should acquire at least 2,500 Chinese characters by the end of the second grade. In addition, 110 students in grades 3–4 did not reach level 1, and 145 students in grades 5–6 did not reach level 2, indicating that about 5.1% of the students failed to achieve their curriculum benchmark on time.

Discussion

As one of the most widely spoken languages in the world, Chinese has seen growing interest in research and evaluation. For the first time, we developed a Chinese character size test (CCST) for Mandarin Chinese primary students through rigorous procedures, following the Chinese curriculum criterion, the unified textbooks, and Messick's validity framework. We provided solid and comprehensive psychometric evidence on the item bank quality, test reliability, inferential score reliability, criterion-related validity, and empirical validity of the CCST. In addition, data from a representative sample of 7,459 students were analyzed to construct the criterion-referenced and norm-referenced character size scores, which are absent in a substantial proportion of existing CR tests. Results showed that the CCST is a convenient, reliable, highly interpretable, and open-access test with solid psychometric quality and comprehensive scoring feedback. Thus, this work enables teachers, researchers, and stakeholders to evaluate the Chinese character size of students whenever applicable.

Extensive meta-analytical studies have demonstrated that feedback is one of the most powerful influences on learning, and effective developmental feedback answers questions about "Where am I going?", "How am I going?", and "Where to next?" (Hattie & Timperley, 2007). In response to the modern demands of "assessment for, as, and of learning," the present study extends beyond the satisfactory psychometric properties provided by traditional CR tests (Hamada et al., 2021; Wen et al., 2015) by providing the CCST with two additional and novel functions: interpretable scores and growth measurement. The interpretable character size scores provide clear diagnostic feedback that is aligned with the curriculum criterion and classroom instruction, supporting students' self-reflection and goal-setting for progress. This is achieved through top-down test design, the curriculum-standards-aligned item pool, and the application of the item response theory framework. Results also indicate that the CCST features a convenient item format suitable for large-scale applications, demonstrating high reliability across the CR developmental continuum and exhibiting strong discrimination power across children of different grade levels, regions, and genders. Therefore, the current CCST enables direct tracking of students'growth across grade levels, and teachers and researchers can leverage the publicly available, large-scale normative data to extend the test into a reusable formative assessment tool that is applicable by semester or learning stage. This tool can effectively showcase and motivate students' progress, thus reinforcing long-term learning outcomes.

Importantly, the validation of the CCST item bank has yielded promising results in terms of item quality and representativeness. The item bank is mapped to the 3,500 commonly used Chinese characters from the national Chinese curriculum criterion, which overlaps with the *List of Commonly used Chinese Characters in Modern Chinese* (1988

edition) by 97.1%, covering a large portion of the most commonly used Chinese characters. This makes the item bank useful to a wide range of researchers and practitioners, as it allows for flexible and tailored assessments within the IRT framework. For example, if the goal is to compare the instructional effects of different primary schools, one would select items that optimize measurement precision across the continuum. Alternatively, a study of whether students have met predetermined standards may benefit from items that provide maximum precision in specific ranges that are close to the educational benchmarks (e.g., selecting the most discriminating items within a certain range of item difficulty). Furthermore, by dynamically supplementing and adjusting the character pool (e.g., adding low-frequency or specialized characters), the CCST can be expanded to meet the needs of different age groups and language contexts, such as middle school students, learners of Chinese as a second language, or adults, serving as a precise assessment tool for their character size and literacy. Because the item bank has been calibrated to the same latent scale, the test efficiency and accuracy of the CCST will be greatly enhanced in the future by incorporating recent methodological advances in computerized adaptive testing (Zhang & Chang, 2016). This would allow the CCST to have broad utility for research and in-depth educational analysis, particularly in terms of its sensitivity to change over time in response to teaching and learning.

The observed criterion-related validity patterns merit particular attention. Our finding of moderate-to-strong correlations between character size and reading ability is consistent with coefficients reported in previous studies using character recognition measures (McBride-Chang & Chen, 2003; Pan et al., 2021). The gradual attenuation of these correlations across grades mirrors the developmental trajectory predicted by the simple view of reading framework, wherein decoding skills become automatized and contribute proportionally less to reading comprehension in higher grades (Catts et al., 2005; Cutting & Scarborough, 2006). Notably, the CCST demonstrates stronger predictive validity for academic performance ($r = 0.43$–0.71) than other measures like the web-based Chinese Character Recognition Assessment (Tseng et al., 2016; $r = 0.35$–0.56). This enhanced ecological validity likely stems from our stratified sampling approach that balanced high-frequency characters from Chinese textbooks and adult corpus.

To the best of our knowledge, the CCST provides the first norm-referenced and criterion-referenced view of character size development in children in grades 1–6. The norms provide benchmarks for typical language development and may help to identify early on children who are significantly behind their peers in character size and growth. For example, as with vocabulary size in alphabetic languages (Biemiller, 2005; Duff & Brydon, 2020), the rate of growth in character size slows over time, as indicated by smaller annual increases. This information may provide a more holistic picture of a child's language development and enable the detection of early signs when a child shows a different trend from expectations relative to his or her peers. In addition, the proportion of primary students in grades 3–6 who have not met their curriculum benchmarks is approximately 5.1%, suggesting that most Chinese children can achieve the goals of Chinese character recognition after several years of basic education. This ratio is also consistent with previous studies that found the prevalence rate of Chinese children with developmental dyslexia to be 5% in Shantou City (Lin et al., 2020) and 4.9% in Guangzhou City (Cai et al., 2020). These similar ratios raise an interesting question about the potential use of the CCST in identifying children with dyslexia. Overall, the norms can serve as a useful tool for educators, speech–language pathologists, and other professionals who work with struggling children to support their language development and improve their reading skills. The CCST could be beneficial for identifying struggling children at an earlier moment, guiding teachers in providing appropriate instruction (Wright & Cervetti, 2017), and even serving as an indicator of education quality.

The present study shows that the CCST is an accurate and valid tool for assessing character size in Mandarin Chinese primary students. However, this research is only a start. In particular, this study primarily measured students'mastery of 3,500 commonly used Chinese characters, but the fact is that children can learn far more diverse Chinese characters than what is available in textbooks. As the study shows, in addition to classroom instruction, incidental learning while reading is also an important way for children to gain vocabulary (Ku & Anderson, 2001). Therefore, the actual character sizes mastered by high-grade students may be larger than estimated, and caution should be taken when interpreting their scores. Future studies are needed to verify the empirical reliability of the proposed character size scores and to supplement and assemble more balanced items, especially those targeting less commonly used characters for higher-level students. In addition, although the test design controls for word stem frequency and homophone interference to make students'recognition performance independent of word knowledge, character–word-level interactions may still partially influence the interpretation of results. Characters with more meanings and compound word forms may appear easier to recognize due to the assistance of word context. The CCST may reflect not only the cognitive processing ability for "form-sound-meaning connections" but also students'whole-word processing knowledge (Hoosain, 1991). Future research could further clarify the specific mechanisms of character–word interactions through experimental designs (e.g., comparing recognition performance for isolated characters versus characters embedded in words). The public

availability of the CCST data allows researchers to investigate these cognitive and psycholinguistic factors more thoroughly, offering opportunities to examine how item features influence recognition accuracy, as well as how different distractors may contribute to errors. Future studies could explore these dimensions through more granular analyses of item-level performance, providing a deeper understanding of the factors that shape character recognition. This line of research could help refine the CCST and enhance its utility as both a diagnostic and developmental tool in educational settings.

Furthermore, we did not account for the impact of tonal variations in dialects when designing the items. This is because, in recent years, the Ministry of Education of China has promoted standardized Mandarin pronunciation and written language teaching nationwide, and we expect CCST results to align with national curriculum standards and teaching evaluation requirements. Despite our success in obtaining a large and demographically diverse sample in terms of grade, gender, and districts, our sample is limited to Beijing, one of the origin locations for standard Mandarin. It is possible that our normative sample differs from the general population in China since we did not include children from dialect areas and poorer areas, and we should be cautious when generalizing and interpreting the results in areas where dialects are more prevalent. It is foreseeable that with follow-up research in more provinces and regions across China, the CCST could offer a more comprehensive and systematic portrayal of regional differences and the urban–rural achievement gap. As outlined in the Strategic Plan for Developing China into a Leading Education Nation (2024–2035), policymakers and practitioners share a sustained commitment to establishing an "equitable and high-quality basic education system." Future studies could leverage the CCST to explore the potential mechanisms through which educational resources and practices influence regional character size outcomes, further clarifying and addressing the influence of important factors such as sociodemographic characteristics, family resources, and school education.

Conclusion

To date, the Chinese character size test (CCST) is the first open-access test measuring the character size of Mandarin Chinese students in grades 1–6. We applied rigorous test development processes, the item response theory framework, and a representative large-scale sample to construct the CCST and its criterion-referenced and norm-referenced character size scores. Solid and comprehensive psychometric evidence is provided in terms of the item bank quality, test reliability, inferential score reliability, criterion-related validity, and empirical validity. The norm results showed that the average character sizes of primary school students in grades 1–6 were 1,227, 1,898, 2,422, 2,722, 2,932, and 3,060 characters, respectively, and approximately 5.1% of students in grades 3–6 had not met their educational benchmarks on time, indicating that most students experience satisfying development in character recognition. Overall, the CCST could be useful for identifying struggling children at an earlier moment, guiding teachers in providing appropriate instruction (Wright & Cervetti, 2017), and even serving as a direct indicator of education quality.

As different stakeholder groups may be interested in the test, we have offered plentiful supporting materials. The CCST test program, test materials, item bank statistics, and normative tables are all available for use at https://osf.io/ktf5c/. Notably, the test program could also be revised and interpreted in line with the practitioner's intentions. This allows users to experimentally measure and calculate character size in an offline computer or online web Inquisit-based setting.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-025-02680-9.

Acknowledgements We acknowledge all participating teachers, students, and parents. The authors are grateful to Miss Shan Huang, Miaomiao Zhen, and Yan Li for their work during test development and administration. We want to thank Dr. Miaomiao Liu for her advice and support which enhanced the quality of this work.

Authors' contributions YL conceived and conducted the study and wrote and revised the article. WY developed the items and conducted the pilots. HL supervised the study.

Funding This work was supported by grants from the Shaanxi 2023 Teacher Development Research Program Special Project (2023 JSQ015), China Postdoctoral Science Foundation (2024M761899), and Shaanxi Province Postdoctoral Science Foundation to YL, as well as grants from the National Language Commission of the People's Republic of China (WT45 - 41) to HL.

Data availability All data including the pilots, the formal test, the final item bank and the normative tables for test scoring interpretation are all provided at https://osf.io/ktf5c/. The test programs are also provided.

Code availability Analysis scripts are publicly accessible for research and teaching use on the Open Science Framework at https://osf.io/ktf5c/.

Declarations

Ethics approval The study was approved by the Institutional Review Board of the Beijing Normal University.

Consent to participation and publication Parents, guardians, and schools provided consent to the data collection and publication procedures. Only participants with valid informed consent forms were allowed to participate in the study.

Conflicts of interest/Competing interests The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Biemiller, A. (2005). Size and Sequence in Vocabulary Development: Implications for Choosing Words for Primary Grade Vocabulary Instruction. In *Teaching and learning vocabulary: Bringing research to practice* (pp. 223–242). Lawrence Erlbaum Associates Publishers.

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics, 37*(6), 1461–1476. https://doi.org/10.1017/S0142716416000060

Cai, L., Chen, Y., Hu, X., Guo, Y., Zhao, X., Sun, T., Wu, Y., & Li, X. (2020). An Epidemiological Study of Chinese Children with Developmental Dyslexia. *Journal of Developmental and Behavioral Pediatrics: JDBP, 41*(3), 203–211. https://doi.org/10.1097/DBP.0000000000000751

Cai, Z. G., Huang, S., Xu, Z., & Zhao, N. (2022). Objective ages of acquisition for 3300+ simplified Chinese characters. *Behavior Research Methods, 54*(1), 311–323. https://doi.org/10.3758/s13428-021-01626-1

Catts, H. W. (2018). The Simple View of Reading: Advancements and False Impressions. *Remedial and Special Education, 39*(5), 317–323. https://doi.org/10.1177/0741932518767563

Catts, H. W., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities. In *The connections between language and reading disabilities* (pp. 25–40). Lawrence Erlbaum Associates Publishers.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Chang, Y.-N., Hsu, C.-H., Tsai, J.-L., Chen, C.-L., & Lee, C.-Y. (2016). A psycholinguistic database for traditional Chinese character naming. *Behavior Research Methods, 48*(1), 112–122. https://doi.org/10.3758/s13428-014-0559-7

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of Reading Comprehension: Relative Contributions of Word Recognition, Language Proficiency, and Other Cognitive Skills Can Depend on How Comprehension Is Measured. *Scientific Studies of Reading*. https://doi.org/10.1207/s1532799xssr1003_5

Duff, D., & Brydon, M. (2020). Estimates of individual differences in vocabulary size in English: How many words are needed to 'close the vocabulary gap'? *Journal of Research in Reading, 43*(4), 454–481. https://doi.org/10.1111/1467-9817.12322

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory. *Psychology Press*. https://doi.org/10.4324/9781410605269

Guan, C. Q., Fraundorf, S. H., & Perfetti, C. A. (2020). Character and child factors contribute to character recognition development among good and poor Chinese readers from grade 1 to 6. *Annals of Dyslexia, 70*(2), 220–242. https://doi.org/10.1007/s11881-020-00191-0

Hamada, A., Iso, T., Kojima, M., Aizawa, K., Hoshino, Y., Sato, K., Sato, R., Chujo, J., & Yamauchi, Y. (2021). *Development of a Vocabulary Size Test for Japanese EFL Learners Using the New JACET List of 8,000 Basic Words*. *65*.

Hao, M., Shu, H., Xing, A., & Li, P. (2008). Early vocabulary inventory for Mandarin Chinese. *Behavior Research Methods, 40*(3), 728–733. https://doi.org/10.3758/BRM.40.3.728

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hattie, J., & Yates, G. C. R. (2013). *Visible Learning and the Science of How We Learn* (1st ed.). Routledge.

Ho, C.S.-H., Chan, D.W.-O., Lee, S.-H., Tsang, S.-M., & Luan, V. H. (2004). Cognitive profiling and preliminary subtyping in Chinese developmental dyslexia. *Cognition, 91*(1), 43–75. https://doi.org/10.1016/s0010-0277(03)00163-x

Ho, S., Chan, D., Chung, K., Tsang, S., Lee, S., & Cheng, W. Y. R. (2007). *The Hong Kong Test of Specific Learning Difficulties in Reading and Writing for Primary School Students (2nd ed.)*.https://www.semanticscholar.org/paper/The-Hong-Kong-Test-of-Specific-Learning-in-Reading-Ho-Chan/3314f4f09f1c7b5eeec507ddc2192dec1b527b79?sort=relevance&page=3. Accessed 02 Feb 2023.

Hoosain, R. (1991). Psycholinguistic Implications for Linguistic Relativity: A Case Study of Chinese. *Psychology Press*. https://doi.org/10.4324/9780203772522

Hung, L., Wang, C., Chang, Y., & Chen, H. (2008). Development of assessment of Chinese character lists for graders. *Psychological Testing, 55*(3), 489–508.

Ku, Y.-M., & Anderson, R. C. (2001). Chinese Children's Incidental Learning of Word Meanings. *Contemporary Educational Psychology, 26*(2), 249–266. https://doi.org/10.1006/ceps.2000.1060

Leung, M., Cheng-Lai, A., & Kwan, S. (2008). *The Hong Kong graded character naming test*. Centre for Communication Disorders, The University of Hong Kong.

Li, H., Shu, H., McBride-Chang, C., Liu, H., & Peng, H. (2012). Chinese children's character recognition: Visuo-orthographic, phonological processing and morphological skills. *Journal of Research in Reading, 35*(3), 287–307. https://doi.org/10.1111/j.1467-9817.2010.01460.x

Li, H., Wang, X., Liu, M., Fan, Y., & Wu, X. (2022). An analysis and comparison of selected characters in primary school Chinese textbooks' character lists. Curriculum, Teaching Material and Method (in Chinese), 42(06), 104–109. https://doi.org/10.19877/j.cnki.kcjcjf.2022.06.012

Li, L., Yang, Y., Song, M., Fang, S., Zhang, M., Chen, Q., & Cai, Q. (2022b). CCLOWW: A grade-level Chinese children's lexicon of written words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01890-9

Li, Y., Huang, C., & Liu, J. (2023). Diagnosing primary students' reading progression: Is cognitive diagnostic computerized adaptive testing the way forward? *Journal of Educational and Behavioral Statistics, 48*(6), 842–865. https://doi.org/10.3102/10769986231160668

Li, Y., Zhen, M., & Liu, J. (2021). Validating a Reading Assessment Within the Cognitive Diagnostic Assessment Framework: Q-Matrix Construction and Model Comparisons for Different Primary Grades. *Frontiers in Psychology, 12*, 5728–5741. https://doi.org/10.3389/fpsyg.2021.786612

Liao, C.-H., Georgiou, G. K., & Parrila, R. (2008). Rapid naming speed and Chinese character recognition. *Reading and Writing, 21*(3), 231–253. https://doi.org/10.1007/s11145-007-9071-0

Lin, Y., Zhang, X., Huang, Q., Lv, L., Huang, A., Li, A., Wu, K., & Huang, Y. (2020). The Prevalence of Dyslexia in Primary School Children and Their Chinese Literacy Assessment in Shantou, China. *International Journal of Environmental Research and Public Health, 17*(19), 7140. https://doi.org/10.3390/ijerph17197140

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39*, 192–198.

McBride-Chang, C., & Chen, H.-C. (Eds.). (2003). *Reading Development in Chinese Children*. Praeger.

McBride-Chang, C., Lam, F., Lam, C., Chan, B., Fong, C.Y.-C., Wong, T.T.-Y., & Wong, S.W.-L. (2011). Early predictors of dyslexia in Chinese children: Familial history of dyslexia, language delay, and cognitive profiles. *Journal of Child Psychology and*

Psychiatry, 52(2), 204–211. https://doi.org/10.1111/j.1469-7610.2010.02299.x

Messick, S. (1990). *Validity of Test Interpretation and Use*. https://eric.ed.gov/?id=ED395031

Ministry of Education. (2022). *The Chinese language curriculum criterion for compulsory education: 2022 edition.* http://www.gov.cn/zhengce/zhengceku/2022-04/21/5686535/files/6b87c3d3411d45ad9f25f88ee33213b7.pdf

Pan, D. J., Yang, X., Lui, K. F. H., Lo, J. C. M., McBride, C., & Ho, C. S. (2021). Character and word reading in Chinese: Why and how they should be considered uniquely vis-à-vis literacy development. *Contemporary Educational Psychology, 65*, 101961. https://doi.org/10.1016/j.cedpsych.2021.101961

Pan, J., Mcbride-Chang, C., Shu, H., Liu, H., Zhang, Y., & Li, H. (2011). What is in the naming? A 5-year longitudinal study of early rapid naming and phonological sensitivity in relation to subsequent reading skills in both native Chinese and English as a second language. *Journal of Educational Psychology, 103*(4), 897–908. https://doi.org/10.1037/a0024344

Perfetti, C. A., Liu, Y., & Tan, L. H. (2005). The lexical constituency model: Some implications of research on Chinese for general theories of reading. *Psychological Review, 112*(1), 43–59. https://doi.org/10.1037/0033-295X.112.1.43

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching, 53*(1), 109–120. https://doi.org/10.1017/S0261444819000326

Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2022). An R toolbox for score-based measurement invariance tests in IRT models. *Behavior Research Methods, 54*(5), 2101–2113. https://doi.org/10.3758/s13428-021-01689-0

Shen, H. H. (2004). Level of Cognitive Processing: Effects on Character Learning Among Non-Native Learners of Chinese as a Foreign Language. *Language and Education, 18*(2), 167–182. https://doi.org/10.1080/09500780408666873

Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (2003). Properties of School Chinese: Implications for Learning to Read. *Child Development, 74*(1), 27–47. https://doi.org/10.1111/1467-8624.00519

Shu, H., Peng, H., & McBride-Chang, C. (2008). Phonological awareness in young Chinese children. *Developmental Science, 11*(1), 171–181. https://doi.org/10.1111/j.1467-7687.2007.00654.x

Stoeckel, T., Bennett, P., & Mclean, S. (2016). Is "I Don't Know" a Viable Answer Choice on the Vocabulary Size Test? *TESOL Quarterly, 50*(4), 965–975. https://doi.org/10.1002/tesq.325

Su, Y., Li, Y., & Li, H. (2022). Familiarity ratings for 24,325 simplified Chinese words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01878-5

Tseng, C.-C., Chang, L.-Y., Chang, Y.-L., & Chen, H.-C. (2016). Wed-based Chinese Character Recognition Assessment and Its Application on Distance Education of Chinese. *Psychological Testing, 63*(3), 179–202.

Tseng, W.-T. (2013). Validating a Pictorial Vocabulary Size Test via the 3PL-IRT Model. *Vocabulary Learning and Instruction, 2*(1), 64–73. https://doi.org/10.7820/vli.v02.1.tseng

Wang, X. L., & Tao, B. P. (1996). *Chinese Character Recognition Test Battery and Assessment scale for Primary School Children.* Shanghai Education Press.

Wei, T.-Q., Bi, H.-Y., Chen, B.-G., Liu, Y., Weng, X.-C., & Wydell, T. N. (2014). Developmental Changes in the Role of Different Metalinguistic Awareness Skills in Chinese Reading Acquisition from Preschool to Third Grade. *PLoS ONE, 9*(5), e96240. https://doi.org/10.1371/journal.pone.0096240

Wen, H., Li, Y., & Yang, Z. (2020). On the IRT-based literacy quantity inference method. *Applied Linguistics, 01*, 112–120.

Wen, H., Tang, W., & Liu, X. (2015). Design of a test on quantity of literacy for students in the stage of compulsory education. *Applied Linguistics (in Chinese)*, *03*, 88–100. https://doi.org/10.16499/j.cnki.1003-5397.2015.03.010

Wright, T. S., & Cervetti, G. N. (2017). A Systematic Review of the Research on Vocabulary Instruction That Impacts Text Comprehension. *Reading Research Quarterly, 52*(2), 203–226. https://doi.org/10.1002/rrq.163

Yeh, M.K.-C., Gopstein, D., Yan, Y., & Zhuang, Y. (2017). Detecting and comparing brain activity in short program comprehension using EEG. *IEEE Frontiers in Education Conference (FIE), 2017*, 1–5. https://doi.org/10.1109/FIE.2017.8190486

Zeng, W., Huang, F., Yu, L., & Chen, S. (2018). Towards a learning-oriented assessment to improve students' learning—A critical review of literature. *Educational Assessment, Evaluation and Accountability, 30*(3), 211–250. https://doi.org/10.1007/s11092-018-9281-9

Zhan, W., Guo, R., Chang, B., Chen, Y., & Chen, L. (2019). The building of the CCL corpus: Its design and implementation. *Corpus Linguistics (in Chinese), 6*(1), 71–86.

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. https://doi.org/10.3389/fpsyg.2022.753913

Zhang, S., & Chang, H. H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning, 1*(1), 67. https://doi.org/10.1504/IJSMARTTL.2016.078162

Zhang, X., Hu, B. Y., Ren, L., & Fan, X. (2018). Sources of individual differences in young Chinese children's reading and mathematics skill: A longitudinal study. *Journal of School Psychology, 71*, 122–137. https://doi.org/10.1016/j.jsp.2018.10.008