Test Review: The New HSK

Yunwen Su[1] , Sun-Young Shin[2]

Abstract
This review discusses the written sections of the New Hanyu Shuiping Kaoshi (the New Chinese Proficiency Test, commonly known as the New HSK), a standardized test developed to measure the proficiency level of non-native speakers of standard Chinese. The test is administered and published by the Office of Chinese Language Council International under the Higher Education Department, Ministry of Education, People's Republic of China, and has the largest number of test takers among all the tests of Chinese as a foreign/second language. The review gives a brief introduction to the history and characteristics of the test, reports on recent research about its reliability, validity, and fairness, and evaluates its strengths and weaknesses.

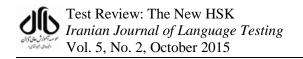*Keywords: The New HSK, Chinese as a second language, standardized test*

1. Introduction

1.1 Test Purpose

The Chinese Proficiency Test (also known as the *Hanyu Shuiping Kaoshi*, hereafter the HSK) is a standardized criterion-referenced test developed to measure the proficiency level of non-native speakers of standard Chinese, or *Putonghua* (普通话). *Putonghua* was established as the official language of China by the People's Republic of China in 1995. It is characterized by the pronunciation of the Beijing dialect, the vocabulary of northern Mandarin, and the grammar of modern vernacular literature (Huang & Liao, 2002). The old HSK was revised in 2009 (henceforth the New HSK) and now includes both written and spoken components. By the end of 2012, a total of 532,909 people (130,700 in China and 402,209 overseas) had taken the New HSK (Chinese Testing International, 2013). This review discusses the written sections of the test. Test scores of the New HSK are used to assist with the college admission for non-native speakers of Chinese. Individual universities can decide their own cut-off scores of the New HSK for admission purposes (Ministry of Education, 2010). Using the 6-level HSK scale, top-tier

---

1 Department of Second Language Studies, Indiana University Bloomington (Corresponding Author), Email: yunwsu@indiana.edu
2 Department of Second Language Studies, Indiana University Bloomington, Email: shin36@indiana.edu

universities in China such as Peking University and Tsinghua University require a 210 or higher

1.2 Administration

The New HSK is administered and published by the Office of Chinese Language Council International (commonly known as Hanban 汉办)[3] under the Higher Education Department, Ministry of Education, People's Republic of China. The test is held every month in official test centers (88 in China and 204 in other countries) and at Confucius Institutes worldwide. Test takers are required to register four weeks in advance for paper-based tests and ten days in advance for internet-based tests. The New HSK scores are available one month after the test date. Individual test takers may obtain HSK certificates and score reports either in person or by mail from test centers. The certificates do not have an expiration date, whereas score reports for university admission purposes are valid only for two years (Hanban, 2010).

1.3 Author and contact information

Chinese Testing International Co. Ltd. (also known as *Hankao Guoji* 汉考国际; hereafter CTI), Room 1803 18/F, Desheng International Center Building B tower, 83 Deshengmenwai Street, Xicheng District, Beijing, China 100088. Test registration: www.chinesetest.cn; general test information and test dates: www.hanban.edu.cn/tests.

1.4 Price

The prices for the written tests are: $20 for Level 1, $30 for Level 2, $40 for Level 3, $50 for Level 4, $60 for Level 5, and $70 for Level 6.

2. Review of Literature

2.1 Test History

Development of the test dates back to 1984. The old HSK put an emphasis on the receptive language skills of reading and listening, and had three different bands—elementary, intermediate, and advanced (Xie, 2011). Hanban discovered general concern over the difficulty of the old HSK in an extensive survey of the test and its administration. The 3000-character vocabulary threshold for the elementary level test was generally felt to be too high a standard for beginning learners by both learners themselves and the instructors (J. Zhang, Xie, Wang, Li, & T. Zhang, 2010).

The new HSK was informed by communicative language testing theories and *Chinese language proficiency scales for speakers of other languages* (Hanban, 2007; hereinafter *The Scales)*. Inspired by *Common European framework of reference for languages: Learning, teaching, assessment* (Council of Europe, 2001; hereafter CEFR), the principles of *The Scales* were developed to encourage learners' communicative competence in using Chinese as a foreign

---

3 *Hanban [The Office of Chinese Council International] is also Confucius Institute Headquarters. China started to establish non-profit public institutions to promote Chinese language and culture in other countries in 2004, which were given the name the Confucius Institute (Hanban, 2013). Their publications may fall under Hanban or Hanban/Confucius Institute Headquarters. The present review sticks to Hanban in its in-text citations for consistency.*

or second language. It differentiates between understanding and production in describing learners' language skills, and provides a five-band all-round description of test takers' ability to use Chinese for communication and exemplar tasks that learners of each band are expected to accomplish (Hanban, 2007).

The New HSK was first officially launched in 2009 (J. Zhang et al., 2010). Table 1 presents the correspondence between the New HSK, *The Scales*, and CEFR reported by Hanban (2010). However, as Jinjun Zhang, Qiu, and Jie Zhang (2009) found in their study, *The Scales* cover low and intermediate levels of standards for Chinese language proficiency, but lack a band that corresponds to proficient users in CEFR.

Table 1. Corresponding levels of the New HSK, The Scales, and CEFR (Hanban, 2010)

Level	Vocabulary	The Scales	CEFR
6	>5000	5	C2
5	2500	5	C1
4	1200	4	B2
3	600	3	B1
2	300	2	A2
1	150	1	A1

2.2 General Description of the Test

Tests of different levels follow different time schedules due to variations in structure (see Table 2). For example, there is no measurement of writing ability at lower levels (Levels 1 and 2), and time allotted for transferring answers from test papers to answer sheets in the listening section varies from level to level. The listening, reading, and writing components of the New HSK are each worth 100 points so the maximum possible score for Levels 1 and 2 is 200 and 300 for Levels 3 to 6.

Table 2. Overall structure and time allotment of the New HSK (based on Hanban, 2010)

Level	Listening			Reading		Writing		Total Time[a]
	Time	Time for Transferring Answers	*N* of Items	Time	*N* of Items	Time	*N* of Items	
1	15 min	3 min	20	17 min	20	/	/	40 min
2	25 min	3 min	35	22 min	25	/	/	55 min
3	35 min	5 min	40	30 min	30	15 min	10	90 min
4	30 min	5 min	45	40 min	40	25 min	15	105 min
5	30 min	5 min	45	45 min	45	40 min	10	125 min
6	35 min	5 min	50	50 min	50	45 min	1	140 min

[a] There is five minutes allotted for filling in personal information at the end of the test.

Tests of different levels carry various task characteristics and should be discussed separately. For the convenience of comparison, tests are grouped so that Levels 1 to 3 are discussed together, then Levels 4 and 5, and finally Level 6. Task characteristics of the tests are summarized from the sample test papers in the official guide (Hanban/Confucius Institute Headquarters, 2010).

3. HSK New Version

3.1 The New HSK Levels 1 to 3

Table 3 presents the stimuli and task types in the New HSK Levels 1 to 3. In general, tests of Levels 1 to 3 in the New HSK target test takers' language use at the sentence level. Specifically, Listening assesses test takers' ability to understand sentences and short conversations, and Reading measures their basic vocabulary knowledge and factual understanding of individual sentences. Language input in both sections in tests of higher levels features longer and more complicated sentences. A writing section is introduced in Level 3 and is called *shuxie* (书写) "spelling and composing", instead of *xiezuo* (写作) "composing". At Level 3, the writing section is not designed to assess test takers' ability to write a passage. It consists of two tasks. Task 1 assesses test takers' ability to write Chinese characters and to construct sentences by sequencing given words. Task 2 assesses test takers' ability to write the corresponding Chinese character according to the *pinyin* that has been given out (Table 2).

Table 3. *Task characteristics of the New HSK Levels 1-3 (based on Hanban, 2010)*

Test	Input	Length of Input[a]	Format	*N* of Items
Level 1				
Listening	5 phrases	2-4 C	T/F (picture)	5
(93 wpm[b])	5 sentences	<12 C	MCQ (picture)	5
	5 conversations	2T	Matching (picture)	5
	5 sentences	<12 C	MCQ	5
				20
Reading	5 words	1-3 C	T/F (picture)	5
	5 sentences	<10 C	Cloze	5
	5 sentences + 5 phrases	<10 C	T/F	5
	3 sentences + 2 conversations	<15 C; 2 T	Matching	10
				25
Level 2				
Listening	10 sentences	<12 C	T/F (picture)	10

(130 wpm[b])	10 conversations	2 T	Matching (picture)	10
	10 conversations	2 T	MCQ	10
	5 conversations	4-5 T	MCQ	5
				35
Reading	5 sentences	<15 C	Matching (picture)	5
	5 sentences	<20 C	Cloze	5
	5 sentences	<30 C	T/F	5
	20 sentences	<20 C	Matching	10
				25
Level 3				
Listening	10 conversations	2 T	Matching (picture)	10
(146 wpm[b])	10 statements	30-50 C	T/F	10
	10 conversations	2 T	MCQ	10
	10 conversations	4-5 T	MCQ	10
				40
Reading	20 sentences	<20 C	Matching	10
	5 sentences + 5 conversations	<20 C; 2 T	Cloze	10
	10 short paragraphs	50-70 C	MCQ	10
				30
Writing	Sentence construction with given words (4-7 words per sentence)			5
	Filling in the missing character (*pinyin* provided) in sentences			5
				10

[a] The length is given in characters (C) for sentences, paragraphs, or passages, and turns (T) for conversations.
    [b] The speed of listening is given in words per minute (wpm) (from J. Zhang et al., 2012).

As the table shows, both listening and reading are tested via a variety of methods, ranging from multiple-choice question (MCQ), true or false statement (T/F), and cloze, to more innovative ones like picture-sentence matching. The variety of tasks employed and the extensive use of pictures make possible the assessment of beginning and low-intermediate test takers' language use.

Other efforts have also been made to accommodate tests of lower levels. Readings in Levels 1 and 2 are presented both in Chinese characters and in *pinyin*, the Chinese alphabet. Listening materials for Levels 1 to 3 are played twice. The speech rate of the language samples is also on the lower end (Level 1: 93 wpm; Level 2: 130 wpm; Level 3: 146 wpm).

3.2 The New HSK Levels 4 and 5
The stimuli and task types in the New HSK Levels 4 and 5 are summarized in Table 4. Language input in listening and reading sections goes beyond sentence level to conversations of multiple turns and paragraphs. MCQ is the dominant format in both sections, supplemented by other tasks such as T/F in listening, and cloze as well as sentence ordering in reading.

Table 4. Task characteristics of the New HSK Levels 4-5 (based on Hanban, 2010)

Test	Input	Length of Input[a]	Format	N of Items
Level 4				
Listening (153 wpm[b])	10 statements	30-50 C	T/F	10
	15 conversations		MC	15
	10 conversations + 5 paragraphs	2 T 4-5 T; 70-100 C	Q MC Q	20
				45
Reading	5 sentences + 5 conversations	<30 C; 2 T	Cloze	10
	10 sets of sentences	3 S/Set	Ordering	10
	14 short paragraphs + 6 long paragraphs	50-70 C; 100-150 C	MC Q	20
				40
Writing	Sentence construction with given words (4-7 words per sentence)			10
	Cued sentence writing (1 picture and 1 word per sentence)			5
				15
Level 5				
Listening (166 wpm[b])	20 conversations	2 T	MC Q	20
	10 conversations + 6 conversations/passages	4-5 T; 120-280 C	MC Q	25
				45
Reading	4 passages	150-300 C	Cloze	15
	10 long paragraphs	100-150 C	MC Q	10
	5 passages	250	MC	2

		-600 C	Q	**0**
				4
				5
Writing	Sentence construction with given words (4-7 words per sentence)			**8** **2**
	Cued essay writing (around 80-character long; cues: 1 picture and 5 words per essay)			**1**
				0

[a] The length is given in characters (C) for sentences, paragraphs, or passages, turns (T) for conversations, and sentences per set (S/Set) for ordering sentences.

[b] The speed of listening is given in words per minute (wpm) (from J. Zhang et al., 2012).

Assessment of receptive language use at these two levels also goes beyond factual understanding. To be specific, the listening section assesses test takers' ability to make inferences based on given information in short statements or conversations (two sentences or turns) and to locate details or facts in longer texts. The reading section measures understanding of text coherence and reading skills such as skimming, scanning, and making inferences. In addition, the writing section assesses test takers' ability to organize their ideas and put them into coherent sentences and paragraphs —at sentence level (cued sentence writing) in the Level 4 test and paragraph level (cued essay writing; two essays of paragraph length) in the Level 5 test.

3.3 The New HSK Level 6

Table 5 presents task characteristics of the New HSK Level 6. All of the three subtests involve language input of much greater complexity than Level 5. The listening section features a rapid speech rate at 234 wpm (cf. 166 wpm in Level 5) and involves stimuli of various lengths from short paragraphs to full interviews and passages. The reading section includes tasks that measure test takers' advanced-level grammatical knowledge (Grammar MCQ) and textual knowledge (Cloze and Sentence Cloze) in addition to the traditional MCQs for passage reading. The writing section is an integrated task, which requires test takers to read a 1000-character narrative in 10 minutes and then to write a 400-character summary of it in 35 minutes.

Table 5. *Task characteristics of the New HSK Level 6 (based on Hanban, 2010)*

Test	Input	Length of Input[a]	Format	*N* of Items
Level 6				
Listening (234 wpm[b])	15 paragraphs	70-100 C	MCQ	15
	3 interviews	600-800 C	MCQ	15
	6 passages	200-500 C	MCQ	20
				50
Reading	10 sets of sentences	4 S/Set	Grammar MCQ[c]	10
	10 paragraphs	50-150 C	Cloze	10

2 passages	400-500 C	Sentence Cloze[d]	10
5 passages	450-850 C	MCQ	20
			50
Writing	A 400-character summary of a 1000-character narrative		1
			1

[a] The length is given in characters (C) for sentences, paragraphs, or passages, turns (T) for conversations, and sentences per set (S/Set) for Grammar MCQs.

[b] The speed of listening is given in words per minute (wpm) (from J. Zhang et al., 2012).

[c] A Grammar MCQs asks the test takers to choose from a set of four sentences the one with grammar error.

[d] A Sentence Cloze asks the test takers to fill in five blanks in a passage by choosing from a set of five sentences.

4. Rating and Scoring

No official explanation of scoring and rating procedures and guidelines has been released for the New HSK. According to *The New HSK rating explanation (self-evaluation)* (CTI, 2012), scoring of reading and listening is straightforward: 0 for incorrect or missing responses and 1 for correct responses to each question. For the writing subtest, questions of different subsections carry different point values (see Table 6). Guidelines for evaluating different writing tasks are provided but the rating rubrics are not available. Rating of sentence construction with words provided focuses on spelling and word order. Evaluation of picture- and word-cued sentence writing tasks is based on grammar, spelling, and content relevance. The four components for evaluating essay writing are content relevance, coherence and logic, grammar, and vocabulary (including spelling).

Table 6. Rating system for the New HSK Writing (CTI, 2012)

Level 3	N of Items	Point Value	Total
sentence construction	5	12	60
character writing	5	8	40
	10		**100**
Level 4	N of Items	Points Value	Total
sentence construction	10	6	60
sentence writing	5	8	4

	N of Items	Points Value	Total
			0
	15		**1**
			00

Level 5	N of Items	Points Value	Total
sentence construction	8	5	40
essay writing	2	30	60
	10		**1**
			00

Level 6	N of Items	Points Value	Total
summary writing	**1**	100	1
			00

The New HSK adopts the mean equating procedure in scoring—raw scores in each section are calibrated with reference to the score distribution in the past three years. The calibrated scores are then converted to percentile scores by linear transformation (J. Zhang, Huang, T. Zhang, Fu, & Huang, in press). Regardless of the level and number of items, each subtest carries equal weight across all the six levels. Starting in late 2013, criterion-referenced scores will be reported with band interpretations (Table 7) and norm-referenced scores against a normative sample established on statistics from the past three years (Q. Zhang, J. Zhang, T. Zhang, Fu, & Huang, submitted).

Table 7. Band interpretation for the New HSK (translated from Hanban, 2010)

Level	Achievement
1	can understand and use some simple Chinese characters and sentences to communicate; ready for continuing their Chinese studies
2	can use Chinese to communicate familiar daily life topics in a direct and simple fashion; high-elementary proficiency
3	can use Chinese to accomplish basic communicative tasks in life, study, and work; can handle most communicative tasks during a tour in China
4	can discuss a relatively wide range of topics in Chinese; can communicate with native Chinese speakers in a relatively fluent way
5	can read Chinese newspapers and magazines, watch Chinese films; can deliver a speech in Chinese
6	can easily understand any information communicated in Chinese; can smoothly express themselves in written or oral form

5. Appraisal of the Test

We evaluated the new HSK test based on Bachman and Palmer's (1996) model of test usefulness that describes key qualities of language tests: reliability, construct validity, and authenticity, and Kunnan's (2014) test fairness model in which fairness is viewed as a test quality that includes absences of bias, access to the test, and test consequences among others.

5.1 Reliability
Reliability is consistency of measurement. In their test usefulness model, Bachman and Palmer (1996) define reliability operationally as consistency across "different sets of test task characteristics" (p.20). Luo et al. (2011) reported reliability estimates (based on Cronbach α) of the New HSK based on item-level scores from Levels 1, 3, and 5 listening and reading sections, and the entire test. As Table 8 shows, reliability indexes obtained from each section and the entire test were acceptable ranging from .85 to .95.

Table 8. Reliability and standard error of the New HSK Levels 1, 3, and 5 (Luo et al., 2011)

Level	α			S_e		
	Listening	Reading	Entire Test	Listening	Reading	Entire Test
1	.72~.88	.80~.92	.85~.95	1.544	1.321	2.069
3	.83~.91	.88~.94	.90~.95	2.204	1.974	3.020
5	.84~.92	.85~.93	.90~.95	2.497	2.762	3.758

5.2 Construct validity
Construct validity, according to Bachman and Palmer (1996), is related to the meaningfulness and appropriateness of the interpretation of test scores. Fu, Zhang, Li, Li, and Zhang (2013) explored the validity of the New HSK Level 5, which had the largest number of test takers among all the six levels, based on statistics from Level 5 tests in April, 2012 with a total number of 6,281 test takers. Correlation analysis indicated moderate correlation between subtests (correlation coefficient: listening and reading= .70, listening and writing= .65, reading and writing= .70) and strong correlation between subtest score and total score (correlation coefficient: listening and total= .87, reading and total= .91, writing and total= .90). Similarly, their exploratory factor analysis results based on the level 5 test scores from 3,141 test takers (random sampling) showed that three factors were extracted, and each subsection of listening, reading, and writing loaded highly onto the corresponding factors representing the three types of language skills. The only exception was Writing Subsection One (sentence construction with words provided, see Table 4) task, which loaded highly on the factor that indexed reading skills. On the whole their results appear to support the construct validity of the test because each subtest is related to each other to some extent, but the subtests seem to tap into different aspects of language skills.

It should also be noted that the tasks utilized by the New HSK reflect its efforts to measure test takers' communicative language ability. It drops the grammar and cloze (multiple choice and fill-in-the-blanks) sections in the old HSK and increases the weight of listening test items (from 30% in the old HSK to 50% to the New HSK). In addition, the New HSK assesses writing at Levels 3 to 6, whereas the old HSK did not measure productive skills until the advanced level. Another improvement is that the New HSK aims to measure test takers' textual competence at higher levels by keeping the multiple choice and cloze of the old HSK in reading sections of Levels 5 and 6 and introducing new task types such as sentence ordering and matching missing sentences in the passage with the four choices.

However, there are still a number of issues that need to be considered to enhance the construct validity of this test. For example, there is a picture-cued essay writing task in Level 5 writing—test takers need to write an essay of about 80 characters by elaborating on a simple picture and bringing into play their imagination and world knowledge. Requiring more than language knowledge to respond to the ambiguous picture may not allow us to make an appropriate interpretation of test takers' writing abilities. In addition, the New HSK needs to broaden the construct of language abilities; there is no evidence that the New HSK measures test-takers' sociolinguistic knowledge, which is important to their survival in the target culture.

5.3 Authenticity
Test authenticity can be demonstrated by comparing the characteristics of the test task with those of target language use (TLU) domains (Bachman & Palmer, 1996). In the New HSK context, the curriculum and materials used in Chinese language programs and classrooms are one of specific TLU domains for the test. In that sense, the New HSK can make the tests more authentic in that they measure a wide range of Chinese language proficiency levels—the Level 1 test targets those who have one semester's Chinese class and the 150 words in their vocabulary pool. Test items and texts are more level appropriate as well. For Levels 1 and 2, pictures are used quite extensively and *pinyin* is provided for all the characters to stimulate test takers' responses. Text length and format change with level, from sentences to paragraphs and short two-turn conversations to multiple-turn conversations, interviews, and passages of various genres (narratives, lectures, and monologic discourse). Test of the writing section is also incremental. As the difficulty of the writing system in Chinese renders measurement of composing ability very hard, sentence construction tasks are included in Levels 3 and 4 and composing tasks in Levels 5 and 6 tests.

Text authenticity in the New HSK improves but still needs to go further. The major problem lies in selection of the listening materials. Some of the passages of Levels 5 and 6 are similar to written texts, with long and formal sentences. Some of the two-turn conversations in the lower-level listening materials have been simplified—probably in consideration of text difficulty and format—that they no longer sound like natural Chinese conversations.

5.4 Fairness
Kunnan (2014) argues that test fairness and justice are important test qualities that should be considered for the appropriateness of score-based decisions and consequences of a test. To

ensure a given test is fair for all test takers, he suggests that a test should be absent of bias, and provide test takers with equal access to the test and its administration conditions, and the learning as well (*the principle of fairness*). He also proposes that an assessment institution should make a positive social impact and advance justice through public reasoning of the assessment (*the principle of justice*).

To date, there has been only one published study regarding the fairness of the New HSK. Fu, Zhang, Li, Zhang, and Xie (2012) evaluated the gender equity of the New HSK Level 6 based on statistics from the total of eight tests in 2011. Results of Differential Item Functioning (DIF) analysis showed that only 3.3% of the 800 test items had item functional differences, and the mean ΔMH was 0.02, with 0 included in the 95% confidence interval. They thus concluded that the test items were balanced between male and female test takers. However, there seems to be no other published DIF studies on other groups including test takers' L1 or academic background.

Nevertheless, there are still a couple of threats to test fairness and justice in the New HSK that need to be overcome. One serious weakness of the New HSK is that it does not provide a screening test to help students decide the appropriate levels of the test to take. It is true that the New HSK provides level-specific vocabulary scale lists, but these lists alone are not sufficient for students to make level decisions. Students who attend the Confucius Institutes can seek guidance from their instructors, but as the New HSK does not position itself as a program-specific test, it should take into consideration the needs of the population not involved in the program to guarantee fair access to the test by different test taker groups.

Another limitation of the New HSK in terms of test fairness and justice is that it is oriented as a general Chinese language ability test that can serve as a reference standard in both academic and professional decision-making procedures (Hanban, 2010), but there is no discussion of the feasibility of the New HSK being used as a one-size-fits-all test. The New HSK needs to clearly state the intended purposes and uses of the test to help test users to make more informed score-based interpretations and decisions about their test takers. More studies on the use of the test and its social impact need to be conducted.

6. Conclusion

The New HSK was designed and developed as a result of people's growing interest in learning Chinese as a foreign/second language and Hanban's intensive efforts to promote Chinese internationally (J. Zhang et al., 2010). Test development has been informed by testing research from the very beginning to ensure the quality of the New HSK. Test-related information, research findings, and test questions are readily accessible to test takers and users so as to help them make informed decisions. Level-specific vocabulary scale lists are available to prospective test-takers and test development follows strict vocabulary guidelines (J. Zhang, 2013). These procedures contribute to promoting the New HSK as the international standardized test of Chinese language ability. However, more research is still needed to provide further evidence for the New HSK's construct- and criterion-related validity, task efficacy, and reliability as a benchmark for Chinese learning. Given that the New HSK test results are increasingly used to

make high-stakes decisions about test takers, the requirements for evidence of such essential test qualities become correspondingly higher.

References

Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Chinese Testing International (2012). *The New HSK rating explanation (self-evaluation)*. Retrieved from http://www.chinesetest.cn/gonewcontent.do?id=5709902

Chinese Testing International (2013). *The New HSK test taker statistics 2013* [internal report].

Fu, H., Zhang, J. [Jinjun], Li, Y., Li, P., & Zhang, T. (2013). The validity research of new HSK Level 5. *Examinations Research,* 2013 (3), 65-69.

Fu, H., Zhang, J. [Jinjun], Li, P., Zhang, T., & Xie, N. (2014) The evaluation of the gender equity of new HSK Level 6. *Examinations Research,* 2014(1). 35-38.

Hanban [The Office of Chinese Council International] (2013). *About us*. Retrieved from http://english.hanban.org/node_7719.htm

Hanban [The Office of Chinese Council International] (2007). *Chinese language proficiency scales for speakers of other languages.* Beijing: Foreign Language Teaching and Research Press.

Hanban [The Office of Chinese Council International]/Confucius Institute Headquarters (2010). *The New HSK test syllabus*. Beijing: The Commercial Press.

Huang, B., & Liao, X. (2002). *Modern Chinese* (3rd ed.). Beijing: Higher Education Press.

Kunnan, A. J. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment: Vol. 3. Evaluation, methodology, and interdisciplinary themes* (pp. 1078-1098). Malden, MA: Willey Blackwell.

Luo, M., Zhang, J. [Jinjun], Xie, O., Huang, H., Xie, N., & Li, Y. (2011). Report on the quality of new Chinese Proficiency Test (HSK). *China Examinations,* 2011 (10). 3-7.

Ministry of Education (2010) Notice about international students' registration with valid New HSK score reports (Jiaowaiting [2010] No. 2). Retrieved from http://www.gov.cn/gzdt/2010-07/13/content_1653154.htm

Xie, X. (2011) Why Hanban developed the new Chinese Proficiency Test (HSK)? *China Examinations,* 2011 (3), 10-13.

Zhang, J. [Jinjun] (2013). On item bank building of new HSK. *China Examinations,* 2013 (4), 21-26.

Zhang, J. [Jinjun], Huang, L., Zhang, T., Fu, H., & Huang, H. (2013). Implementation plan of the mean score equating method in new HSK (Level 6). *Examinations Research,* 2013 (5), 42-47.

Zhang, J. [Jinjun], Li, P., Li, Y., Xie, N., & Huang, L. (2012). New thinking on the New HSK. *China Examinations,* 2012 (2), 50-53.

Zhang, J. [Jinjun], Qiu, N., & Zhang, J. [Jie] (2009). The report of linking research between HSK and Chinese Proficiency Scales for speakers of other languages. *Chinese Examinations,* 2009 (4), 18-23.

Zhang, J. [Jinjun], Xie, N., Wang, S., Li, Y., & Zhang, T. (2010). The report of researching and producing the new HSK. *China Examinations,* 2010 (9), 38-43.

Zhang, Q., Li, Y., Fu, H., & Zhang, J. [Jinjun] (submitted). A comparative study of the paper-based and internet-based New HSK.

Zhang, Q., Zhang, J. [Jinjun], Zhang, T., Fu, H., & Huang, H. (submitted). How to process the score of new HSK (Level 6).