Guidelines and Technical Report for Chinese Literacy Assessments

A collaboration between

STARTALK, Mandarin Institute, Loyola Marymount University and ACTFL

Guidelines and Technical Report for
the Chinese Early Literacy Assessments

One of the challenges that Chinese language teachers face in teaching content vocabulary is that there are very few assessments that can inform instruction and measure vocabulary growth because *prior to the Mandarin Institute Loyola Marymount University 2015 STARTALK Infrastructure program, high frequency words for Chinese L2 learners were not readily defined nor was there a clear way to identify Tier 1, 2, 3 vocabulary.* This resulted in the creation of the K-5 Chinese Word Frequency Dictionary (WFD) for L2 Learners. For our 2016 Infrastructure program, we partnered with ACTFL and developed a collection of contextualized character and vocabulary recognition assessments and one reading assessment by grade to serve as a baseline for future work, that can more accurately track students' vocabulary growth across the curriculum aligned to the WFD, ACTFL, common core and content standards and prove reliability of the K-5 Word Frequency Dictionary. Finally, in the third of the three part series of STARTALK Infrastructure grants, we developed a series of Reading Comprehension Assessments.

This report is organized into two main parts; the first part provides guidance to teachers about how to use, and keep track of student assessment results for character and vocabulary recognition and reading assessments. The second part consists of the technical aspects of the development of assessments to ensure that users understand the reliability and validity of the assessment.

Table of Contents

Overview

Because there are no defined reading levels for materials and no adopted reading assessments in Chinese second language reading for K-5 classrooms, Chinese language programs have had no tools, measure or scales that can assist Chinese language teachers in selecting literature or informational texts appropriate for their students, nor have there existed standardized ways to assess students' reading proficiency. To design literacy instruction that addresses students' reading levels, teachers must first be able to identify their students' language and literacy levels. These Guidelines provide an overview of the sequences of three STARTALK Infrastructure grants provided to the Mandarin Institute and Loyola Marymount University. We review how each of the three phases of funding from STARTALK helped to developed the "building blocks" for Chinese early literacy assessment that lead to the development and validation of character, vocabulary and reading comprehension tests.

The Mandarin Institute-LMU 2015 STARTALK Infrastructure Grant *Building the Prototype of the K-5 Word Frequency Dictionary (WFD)* resulted in the development of the first high frequency word dictionary for L2 Chinese learners which is comprised of ~3,500 words with associated band/grade levels defined.

One of the challenges that Chinese language teachers face in teaching content vocabulary is that there are very few vocabulary assessments that can inform instruction and measure vocabulary growth because prior to our 2015 Infrastructure grant, high frequency words for Chinese L2 learners were not readily defined nor was there a way to identify Tier 1, 2 and 3 vocabulary and thus there are no assessments that directly tie to them. For our 2016 Infrastructure project our team developed a collection of contextualized character and vocabulary recognition assessments that can more accurately track students' vocabulary growth across the curriculum aligned to the WFD, ACTFL and content standards and prove reliability of the K-5 Word Frequency Dictionary.

The second grant (2016), *Building Chinese Early Literacy Assessments for L2 Learners,* built upon the WFD to create and technically validate the following K-5 vocabulary tasks/tests associated with the Word Frequency Dictionary levels:

- Character recognition: 3 tests each with 15 items for a total of 45 items per grade
- Vocabulary recognition: 3 tests each with 15 items for a total of 45 items per grade
- Text-embedded vocabulary assessment: 1 test with 15 items per grade

In addition, for each of these phases, test administration, data collection and scoring guides were developed to serve as references for all participating schools and teachers, especially for those who did not attend the online or onsite training. This manual provides detailed information and guidance on the testing environment and instruction, the data collection tool, data scoring and recording procedures. These are included in the guide and in the appendices.

Collaborators: The Team

With the guidance of an interdisciplinary expert team consisting of Chinese linguists, language, literacy and assessment experts, Master Teachers from partial, dual and full immersion programs and 5 different states representing public, independent and charter schools spanning socio-economic and diverse student populations, and half a dozen research assistants, the team developed a high frequency word list for K-5 Chinese learners and defined word frequency scope for each grade level which can be used as a reference to determine text difficulty of a particular reading material; and character and vocabulary recognition assessments and a baseline reading comprehension assessment.

Beginning in the 2015 Infrastructure award, the program team selected 10 Master Teachers who were nominated by their administrators. Selected teachers came from partial, dual and full immersion programs and have teaching experiences ranging from K-16. 70% of the teachers have a Master's Degree, 50% are credentialed and 90% have completed one or more ACTFL courses which includes OPI and WPT. 30% have been teaching Chinese for over 10 years, 50% for 5-10 years, and 20% for 3-5 years. While Master Teachers varied over the three years, their overall backgrounds and school-type representations remained consistent. Participating Master Teachers were required to attend the one-week long Summer Institutes each year at Loyola Marymount University.

Following is a list of our team which covers the 2015, 2016 and 2017 STARTALK Infrastructure awards:

Program Team

Dr. Michael Everson	Emeritus Associate Professor University of Iowa
H. Yalan King	Executive Director Mandarin Institute

Dr. Magaly Lavadenz	Professor, Dept. Educational Leadership Executive Director, Center for Equity for English Learners Loyola Marymount University
Dr. Ping Liu	Professor of Teacher Education Specialized in Chinese/English Immersion Education California State University Long Beach
Dr. Claudia Ross	Professor of Chinese Department of Modern Languages and Literatures College of the Holy Cross
Paul Sandrock	Director of Education American Council on the Teaching of Foreign Languages
Dr. Helen Shen	Professor, Dept. Asian and Slavic Languages and Literatures University of Iowa
Dr. Carl Swartz	Research Professor, Early Childhood Special Education and Literacy School of Education University of North Carolina, Chapel Hill
Qian Helen Zhou	Ph.D. (candidate), Second Language Acquisition Program University of Maryland Language Assessment Specialist Intern, Mandarin Institute

Master Practitioners

Waisum Buenning	Horizon Elementary Public School	UT	Dual Immersion
Eric Stohl Chipman	North Park Elementary Public School	UT	Dual Immersion
Patty Chung	Lone Peak Elementary Public School	UT	Dual Immersion
Yuching Chung	Washington Yuying Public Charter School	DC	Partial Immersion
Xiu Geng	Chinese American International School	CA	Partial Immersion
Qin Hua	Washington Yuying Public Charter School	DC	Partial Immersion
Shu-Mei Lai	Yinghua Academy Charter Public School	MN	Full Immersion
Hsueh Ting Li	Wedgeworth Elementary Public-School	CA	Dual Immersion
Belinda Liu	Denver Language Charter Public School	CO	Full Immersion
Yinzhu Liu	Chinese American International School	CA	Partial Immersion
Wei Shen	Yu Ming Charter Public School	CA	Dual Immersion
Vera Song	Washington Yu Ying Public Charter School	DC	Dual Immersion
Yuli Sun	Broadway Elementary Public School	CA	Dual Immersion
Xiaohong Sui	Chinese American International School	CA	Partial Immersion

Vivian Wang	Broadway Elementary Public School	CA	Public
Wenjuan Wang	Yinghua Academy Charter Public School	MN	Full Immersion
Haii West	Brigham Young University	UT	Dual Immersion
Hui-Tzu Wu	Yinghua Academy Charter Public School	MN	Full Immersion
Xinyi Xu	Yuming Charter Public School	CA	Dual Immersion
Pei Pei Xue	Stewart Elementary Public School	UT	Dual Immersion
Shanshan Yang	Coronoda Elementary Public School	AZ	Partial Immersion

Research Assistants-(2015-2016)

Jia Jiang	Lead Research Assistant for Data Collection and Input
Bing Guo	
Jia Hu	
Xuanping Li	
Sihong Liu	
Weiqing Liu	
Zicun Zhao	
Shan He	Visiting PhD student at University of Iowa, Beijing Normal University

Part I: Guidelines for Using the WFD and Assessments

This first phase (2015-2016) of the document provides teachers and other users guidance on how to use the Word Frequency Dictionary and in how to administer and keep track of the results of each of each of the different types of assessments developed by our team.

Word Frequency Dictionary

The principal goal of the Mandarin Institute-Loyola Marymount University (LMU) STARTALK Infrastructure Building the K-5 Word Frequency Dictionary for Assessing Early Literacy (2015) is to build the foundation for early Chinese reading instruction by creating lexiled vocabulary levels and assessments.

Given the lack of assessments, measures and tools to support teachers in selecting grade-level reading materials and texts appropriate in a variety of Chinese immersion programs, the first baseline word frequency dictionary for K-5 Chinese L2 learners has been developed. These leveled lists along with an online searchable database tool can be used to determine text difficulty of a particular literature or informational reading materials/texts. Together, the word lists and searchable database can be used to support teachers in the design of literacy instruction that addresses their students' reading levels.

To make the K-5 WFD more useful to classroom instructors, the high frequency words were aligned with the vocabulary requirements of Common Core and Content Standards for each grade level. The word frequency dictionary served as the foundation to creating a formula for measuring text complexity, and thus readability for grade level texts.

The Process

1. We focused our scope on K-5 Chinese immersion programs and on tackling the issue of analyzing text complexity. The most important factor in analyzing text complexity is determining the frequency of words. There is currently no Dictionary of High Frequency Words for Chinese L2 learners (learning Chinese as a second language). Texts that contain a large number of high frequency words will inherently be easier to comprehend than text that contains low frequency words. In order for this listing to be useful to classroom instructors, we needed to align the high frequency words with vocabulary requirements, Common Core and Content Standards for each grade level. These in combination, along with some other factors will serve as the foundation to creating a formula for measuring text complexity.

2. The Leadership Team determined that in order to create the K-5 Word Frequency Dictionary (WFD), we first needed to create and analyze a corpus of at least 2,500,000

Chinese characters in order to be statistically sound. In creating the corpus, we targeted a minimum of 100 books, narrative and informational, per grade levels by common core standards across multiple categories:

- Published children's literature books
- Online academic materials
- Textbooks
- Readers

Our Master Teachers categorized books by subjects aligned with Common Core State Standards and Subjects by grade and classified children's books by genre and grade levels. Teachers then selected and created sample sentences from books aligned with key vocabulary. * We had to use caution so as not to skew the representative sample of materials targeted towards native speakers by comprehending emergent second language readers' literacy vocabulary especially at the kindergarten level. However, this selection of materials is representative of what is currently being used in K-5 Chinese immersion programs in the U.S. Based on these materials, we created a corpus of 2,595,956 characters.

3. We used a statistical software program that was adapted to assist with frequency analysis of our corpus of nearly 2.6 million characters. The analysis was then manually calibrated and broken out into bands/levels. The WFD bands are equivalent to the level of text complexity that will be used to select appropriate texts for Chinese K-5 immersion classrooms, or high frequency dictionary bands aligned with K-5 grade levels.

We consider the development of this word frequency list as a critical first step. The K-5 Chinese Word Frequency Dictionary can serve as a baseline for K-5 immersion curriculum in the U.S. and abroad. When using the list to determine the text difficulty for a particular grade level, we recommend that ~70% of the words in a text should fall within the range of the Word Frequency Dictionary band for a particular level.

Using the K-5 Chinese Word Frequency Dictionary

The K-5 Word Frequency Dictionary for L2 Chinese Language Learners can help to:

- Predict student's reading comprehension by grade
- Measure text complexity – teachers can determine if a reading text is at a certain level
- Provide vocabulary guidelines for textbook writing and reading material selection
- Inform the development of assessments

A band includes a collection of words that occurred most frequently in the corpus of characters for a given grade level. This corpus of characters was an assembly of a variety of text across different subjects for Chinese immersion programs.

K-5 Word Frequency Dictionary Bands

GRADE	WORDS FROM LOWER GRADE(S)	NEW WORDS	TOTAL =WORDS FROM LOWER GRADE(S) +NEW WORDS
K	0	300	300
1	300	400	700
2	300+400=700	500	1200
3	300+400+500=1200	600	1800
4	300+400+500+600=1800	700	2500
5	300+400+500+600+700=2500	849	3349

Question: Do the number of words per band indicate the number of words that are the learning target at each grade level?

Answer: No. The number of words per band do not reflect expectations of student learning at each grade level. However, they can be helpful to teachers as they plan instruction and design assessments.

Question: How did you select which words go in each band level, and how many to include?

Answer: These are the words that occurred frequently for that grade level in the corpus of characters assembled. For Band K, 300 words were found frequently in the corpus of characters for the kindergarten level.

To access the online searchable Word Frequency Dictionary and detailed word lists by grade please go to http://mandarininstitute.org/K-5%20WFD

The WFD can be used as a tool to help Mandarin teachers assess the grade level appropriateness of reading materials. With selected reading text in place, instructional planning

can be designed to support content and academic language development. The application of the WFD in teaching context can include the following steps:

Checklist for Using the K-5 Chinese Word Frequency Dictionary

The following checklist details the steps on how to use the K-5 Word Frequency Dictionary. Follow the provided links to access the needed resources.

✓	STEPS FOR USING THE WFD	RESOURCES
	1. Review a reading passage/story and align the text with grade level appropriate content standards	Common Core, content and ACTFL Proficiency standards
	2. Identify or select content-based key academic vocabulary	Table for Vocabulary Grade Bands and Detailed Word List http://www.mandarininstitute.org/K-5%20WFD
	3. Use the online searchable WFD to check vocabulary (characters and words) frequency by grade	Online searchable WFD http://www.mandarininstitute.org/K-5%20WFD
	4. Plan instructional activities to organize the identified vocabulary to learn concept and language	
	5. Guide students to apply vocabulary in context for content/language development during instruction	
	6. Create opportunities for students to demonstrate learning and understanding through the use of vocabulary in context with aligned formative and summative assessment	Appendix I MI-LMU STARTALK K-5 assessments

PART II: Technical Report for Assessment Development and Testing

Developing the Assessments

During Phase 2, the team created a series of assessments from K-5th grade which included character and vocabulary recognition assessments comprised of 45 items each. The team also created one reading comprehension assessment comprised of 15 items to be used as a baseline for Phase 3, which involved developing a series of model reading comprehension assessments for 3rd through 5th grade Chinese immersion programs.

Master Practitioners

The master teachers were integral in the development of the K-5 Word Frequency Dictionary and the character, vocabulary and reading comprehension assessments. Throughout the 2015, 2016 and 2017 STARTALK Infrastructure grant periods, they participated in each of the week-long Summer Institutes at LMU, learned how to apply the WFD to select grade appropriate reading materials and plan instruction driven by assessment with highlighted vocabulary. They had an opportunity to share ideas, receive training in assessment development, and create testing items collaboratively. Specifically, they participated in the following activities:

- Practiced how to apply the WFD step by step using a content based sample story, which they were encouraged to apply in their own classrooms to build a connection between instruction and assessment
- Reviewed and discussed the AAPPL assessment samples in content and format
- Compiled a character and vocabulary list based on content standards with a focus on science
- Explored how to develop test items on vocabulary by definition, synonyms, antonyms, hypernyms and hyponyms, including characters and pinyin that are specific to Chinese language
- Analyzed summative assessment samples of Chinese as a native and foreign language in format and content
- Created K-5 testing sets through teacher grade level group work with cross-grade discussion to make clarifications as needed
- Implemented AAPPL assessments and the STARTALK Chinese Early Literacy assessment tasks in their classrooms/schools, and collected and recorded the results in order for us to complete the validation studies.

Materials Selection

The primary sources of text or content for the assessment were chosen from materials used in the target grade level classrooms and the compiled K-5 Chinese text corpus word-processed in

2015, in addition to online and other resources. The K-5 text corpus, comprised of approximately 2.6 million Chinese characters, includes different types of text such as textbook excerpts, children's stories and other information by grade. Some of the text selections were revised for the purpose of grade level appropriateness. In addition, a list of content-based vocabulary was compiled in the 2016 summer institute. By adapting Beck, McKweon and Kucan's (2002) concept of tiered vocabulary instruction to second language learners, our Master Practitioners learned to identify the three types of vocabulary students' need to acquire, both language and content in Chinese immersion classrooms. [1] Included in these levels are attributes such as word frequency, complexity and domain-specific academic terms in determining tier levels. We needed the WFD to do this work. The list was used as reference in test development. Finally, the test items were used in the classrooms and schools of the master teachers and were considered and revised as appropriate.

Character, Vocabulary Recognition and Reading Comprehension Assessments

The assessments, in simplified Chinese, are a K-5 test collection to assess character and vocabulary recognition for students of Mandarin immersion programs. Each grade has an assessment set that includes multiple choices of characters, vocabulary and reading comprehension questions. These are included in Appendix I. Pictures, characters, vocabulary, sentences and passages were the basic elements in different types of tests. For each type of test, a sample is provided to help students understand the directions. Along with the test sets, answer keys by grade are included. Each grade set was created to adhere to the following:

- Label grade at the beginning of page 1 for each of the files
- Include references of any selected text in the corpus
- Cite the source of any new text not included in the corpus
- Use "Adapted from…" to note any revised text

The structure and format of the assessment in characters, vocabulary and reading/listening comprehension are summarized as follows:

Assessment of characters in oral pronunciation and recognition

i. Sound out a list of characters
ii. Character and picture match by choosing one from a list of choices. Distractions for characters/vocabulary are those that sound alike, have similar meaning or look alike

[1] Beck, Isabel L., McKeown, Margaret G., and Kucan, Linda. (2002). *Bringing words to life.* New York, NY: The Guilford Press

Assessment of vocabulary identification

i. Choose the vocabulary words for a given picture
ii. Choose a character from a list to form a vocabulary
iii. Choose a vocabulary for a category (content-based).

Reading Comprehension Assessment Tasks

The assessments developed in Phase III are in simplified Chinese and are a Grade 3-5 test collection to assess reading comprehension for students of Mandarin immersion programs. Each grade has an assessment set that includes three content areas: Language Arts, Social Science, and Math. For each content areas, the test items are designed in three formats: multiple choice questions, fill-in-the-blanks, and short answers. These are included in the Appendix. Passages were the basic elements in Language Arts and Social Science questions. Math items mainly consist of individual problems. For each type of test, a sample is provided to help students understand the directions. Along with the test sets, answer keys by grade are included for multiple choice and fill-in-the-blank questions. In addition, grading rubrics are provided for short answer questions. For multiple choice questions, students are expected to answer a comprehension question by choosing one out of four on a list. For fill-in-the-blank question, students are expected to provide a word in Chinese to complete a sentence based on his or her understanding of the question and the reading passages. For short answer question, students are expected to provide at least one sentence or solution that corresponds to the question. Here is an overview of the test items.

Grade 3

	Language Arts	Social Science	Math	Total
Multiple Choice	17 items	17 items	11 items	45 items
Fill-in-the-blank	3 items	4 items	9 items	14 items
Short Answers	10 items	9 items	10 items	29 items
Total	30 items	30 items	30 items	90 items

Grade 4

	Language Arts	Social Science	Math	Total
Multiple Choice	20 items	19 items	16 items	55 items
Fill-in-the-blank	3 items	4 items	7 items	16 items
Short Answers	10 items	12 items	13 items	35 items
Total	33 items	35 items	36 items	104 items

Grade 5

	Language Arts	Social Science	Math	Total
Multiple Choice	20 items	20 items	10 items	50 items
Fill-in-the-blank	N/A	N/A	12 items	12 items
Short Answers	11 items	10 items	7 items	28 items
Total	31 items	30 items	29 items	90 items

Data Collection Procedures

Given that newly developed assessments should be validated before being employed, a large number of students from various schools were invited to take these assessments in order to evaluate the reliability and validity of these tests. Step-by-step guidance was provided to all participating schools and teachers to ensure consistency of the data collection procedures. This section illustrates the detailed data collection processes.

Training

Before the field testing began for each of the phases, online training was given via WebEx to all master teachers involved in the test development and who returned to their classrooms to administer the corresponding sets. At a minimum, at least one teacher from each participating school participated in the online training. The training focused on the number and types of assessments to be tested, how the tests should be administered, how each item was to be scored and how to record all the data. A data collection tool was designed and shared with the master teachers prior to the meeting and details regarding use of the tool was explained during the online training. Master teachers were provided with ample time in the Question and Answer section of the training where all questions were addressed and follow-up communications where any remaining uncertainty was clarified. During the online training, the data collection timeline for each participating school was determined.

Data Collection Timeline- Phases II and III		
Phase II	**Phase III**	**Task**
Early October 2016	Early October 2017	WebEx Test Administration Preparation Meeting
Mid October 2016	October 7th & 14th, 2017	Submit AAPPL registration form
October 2016	November 2017	AAPPL reading test
October – November 2016	October-November 2017	Field Testing of assessments developed at the Institute
December 2016	December 2017	Record data and complete data collection tool

Guidance for Administering and Scoring the Assessments in Phase II

The following section briefly provides recommendations on how best to administer the character and vocabulary recognition, and reading comprehension assessments, as well as how to keep track of student assessment results in order to monitor progress over time.

Checklist for Administering the Assessments

The following checklist details the steps on how to prepare for and administer the Character, Vocabulary and Reading Comprehension assessments.

✓	STEPS FOR ADMINISTERING THE ASSESSMENTS
	1. Test administrators and teachers (if not the same person) should work together to determine the most appropriate testing environment based on the number of students and estimated time needed to complete each test.
	2. The test administration should be conducted in a secure environment.
	3. Establish procedures to maintain a quiet testing environment throughout the test session, recognizing that some students will finish more quickly than others. If students are allowed to leave the testing room when they finish, explain the procedures for leaving without disrupting others. If students are expected to remain in the testing room until the end of the session, instruct them on what activities they may engage in after they finish the test.
	4. Make sure students do not have reference resources including dictionary, internet, books, etc.
	5. To ensure that all students are tested under the same conditions, test administrators must adhere to the instructions and make sure that they instructions are well explained. Lead students through the examples first.
	6. Test administrators should try to maintain a natural classroom atmosphere during the test administration. Before each test begins, the teacher should encourage students to do their best.

Checklist for Scoring the Assessments

The following checklist details the steps on how to score the Character, Vocabulary and Reading Comprehension assessments. Refer to Appendix II for a sample Data Collection Tool.

✓	STEPS FOR SCORING ASSESSMENTS AND RECORDING ASSESSMENT RESULTS
	1. Students receive 1 point for each correct item, and 0 points for each incorrect item.
	2. The maximum score students can get for each test is 15, and the minimum is 0.
	3. Given that all items are multiple choice items, there is only correct answer for each item. If more than one answer or no answer is selected, no point will be given for this item.
	4. Scoring & Data Recording • For all multiple-choice format questions, please record students' answers (A/B/C/D). When there is a missing answer, please record it as "*". Please keep in mind that some fill-in-the-blank or cloze tests are also in multiple choice format. • For all fill-in-the-blank or cloze tests, where only one character or word is required in the answer, please record 1 if students answer it correctly and record 0 if the answer is incorrect or missing. • For all open-ended questions, where at least one sentence is required in the answer, please record 0-4 points or 0-3 points based on Grading Rubrics for each grade. Grading rubrics for each grade can be found in the following tables. • Grade 5 Only: please record both comprehension and writing scores for each MLA and Social Science open-ended questions based on the grading rubrics. • The process of scoring is repeated for all reading comprehension tests.

Onsite training was also performed in some schools such as Yu Ming Charter School where multiple classes participated in the field testing. Face-to-face instruction on test administration and data collection were offered to teachers who did not take the online training at these schools.

Data Collection Tool and Test Administration Guide-Phase II and Phase III

To facilitate the data collection process, a data collection tool was created for participating schools and teachers to record all the test information and results. This tool is comprised of a set of spreadsheets including the student's demographic information (name, school, district, state, teacher name, gender, grade, number of years in the immersion program), the student's ID (a six-digit number combined by school code, test grade, and assignation of a unique number), test information and item-specific scores for each set of assessments. The data collection tool is included in Appendix II. In addition, a Test Administration Guide was developed in order to inform consistent administration of the assessments.

Phase II Participating Schools and Grades

A total of 22 classes (one teacher per class) from five schools participated in the field test, which included one Grade K class, four Grade 1 classes, three Grade 2 classes, six Grade 3 classes, five Grade 4 classes and three Grade 5 classes. Participating schools comprised public, charter, and independent schools from California, Utah, Minnesota, and Washington DC. Below is a synopsis of the test sites by school and grade.

Number of Classes Assessments were Administered in Phase II

SCHOOL	Kinder	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5
CAIS	0	2	2	2	2	2
Draper	1	1	1	1	1	0
Ying Hua	0	1	0	1	0	0
Yu Ming	0	0	0	2	1	1
Yu Ying	0	0	0	0	1	0
Total Classes	**1**	**4**	**3**	**6**	**5**	**3**
	Classes that also took AAPPL assessment					

In addition to the assessments developed at the Institute, several students also took the AAPPL interpretive reading test, which served as a high-stake standard test for our validity testing. Based on advice from ACTFL experts, we decided that only Grade 3 to Grade 5 students should

take the online AAPPL reading test. A total of 12 classes from five schools took the AAPPL reading test.

Part III: Data Analysis and Results

All assessment data were collected from participating teachers between November 2016 and January 2017. Item-level scores of AAPPL reading test results were gained from ACTFL. These data were then cleaned and categorized based on grade level and test type before being evaluated through reliability and validity testing.

Descriptive Data Summary

A total of 731 students registered for the field test, which included 56 Kindergarten students, 133 Grade 1 students, 128 Grade 2 students, 158 Grade 3 students, 171 Grade 4 students and 85 Grade 5 students. The following table depicts the student number across school and grade.

Phase 2-(Character and Vocabulary Knowledge) School Testing Summary

	CAIS	DRAPPER	Ying Hua	Yu Ming	Yu Ying	Total per grade
K		56				56
1	48	56	29			133
2	44	58	26			128
3	52	54		52		158
4	48	53		51	19	171
5	48			37		85
Total per School	240	277	55	140	19	731

Given that some students who registered were not able to take the assessments due to illness or absence, invalid cases were removed from the original data sets by grade and test type to avoid missing data. The following table presents a final summary of valid data by grade and test type, which were submitted for reliability and validity analysis. In addition, 325 students from four schools took the AAPPL interpretive reading test, whose results were then used for concurrent validity testing

Valid Data by Grade and Test Type

	Character Recognition	Word Recognition	Reading Comprehension
K	56	56	56
1	103	103	103
2	119	131	118
3	152	153	152
4	171	171	169
5	79	83	84
Total	683	697	682

Reliability Results

Reliability is the notion that the test produces stable and consistent results over time. Rasch analysis was performed to decide item reliability of each test. Rasch analysis has been applied to assessments in a wide range of disciplines. The Rasch model is the only item response theory model in which the total score across items characterizes a person totally. By conducting Rasch analysis, the total item reliability for each test as well as the fitness/unfitness of specific items in the test were obtained. In this way, poor-quality test items could also be diagnosed. Eighteen Rasch analysis were run for corresponding tests and grades, the results of which can be found in the following table.

Rasch Analysis Results

	Character Recognition	Word Recognition	Reading Comprehension
K	Item reliability 0.95	Item reliability 0.92	Item reliability 0.84
1	Item reliability 0.92	Item reliability 0.93	Item reliability 0.68
2	Item reliability 0.93	Item reliability 0.95	Item reliability 0.84
3	Item reliability 0.93	Item reliability 0.95	Item reliability 0.80
4	Item reliability 0.96	Item reliability 0.94	Item reliability 0.94
5	Item reliability 0.90	Item reliability 0.92	Item reliability 0.92

According to the above results, all of the Character Recognition and Word Recognition item reliability scores were above 0.90, which shows directly that these test items developed at the STARTALK Institute are of very high quality. Specifically, these data demonstrate that the difficulty levels of the items are variant enough to differentiate participants' abilities.

All of the Reading Comprehension scores were still well above .70 except for Grade 1. Item-specific analysis suggests that the first item of Grade 1 Reading Comprehension test was biased, indicating a revision of the item or its distractors. It is noticeable that item reliability scores of Reading Comprehension tests were generally lower than those for Character Recognition tests and Word Recognition tests. The main factor accounting for the difference is the number of items developed for these tests. Given that our main focus was on vocabulary and character knowledge, 45 items were created for Character Recognition as well as Word Recognition tests, whereas only 15 items were created for Reading Comprehension tests. Considering the limited number of items for Reading Comprehension tests, the results can be considered satisfactory and their primary purpose is to be used as a baseline for future reading comprehension assessment development to be conducted in 2017.

Coefficient alpha (i.e., Cronbach's alpha), which is the most popular way to estimate test reliability, was also calculated. It measures the extent to which the items provide consistent information regarding the students' mastery of the domain. Coefficient alpha was calculated based on the number of items on the exam, proportion of examinees who answered each item correctly, and sample variance for the total score. Results were summarized in the following table.

Coefficient Alpha

	Character Recognition	Word Recognition	Reading Comprehension
K	0.622	0.622	0.444
1	0.941	0.923	0.825
2	0.862	0.899	0.828
3	0.926	0.874	0.622
4	0.870	0.917	0.685
5	0.859	0.923	0.796

As can be seen from the table, most of the test reliability indexes are well above .70, indicating high consistency of the test items. Relatively lower test reliability was found in Kindergarten, mainly due to the fact that the number of participants was smallest in Grade K and all of them were from one school, which reduced the sample variances.

Validity Results

We used both content and concurrent validity in order to determine the degree to which there is a match between test questions and the content or subject area they are intended to assess. Content validity answers the question: *Do our assessments measure character recognition, vocabulary recognition and reading comprehension for K-5 Chinese language learners?* Content validity was achieved with Chinese literacy content experts in the following ways:

1. All of our test items were developed by master teachers who are domain experts in the content.
2. Language and literacy domain experts provided training to these master teachers, ensuring that items were selected from the appropriate content.
3. All items were cross-examined by peer master teachers from different schools and grades.

Concurrent validity was measured to assess whether our assessments have strong criterion validity--namely whether our assessments measure Chinese reading abilities. The AAPPL Reading test was utilized as the benchmark test with which to establish concurrent validity. As a result, correlations between AAPPL test scores and corresponding scores in our tests were calculated in order to measure concurrent validity. It is important to note that the AAPPL Interpretive reading test items are not aligned to content standards nor are they grade-level specific. Also, AAPPL interpretative reading test items do not include discrete items such as the character and vocabulary recognition test items we developed here. Thus, there is as expected, variance in our results in using it as our high stakes test. The concurrent validation results using the AAPPL are presented in the following table.

Concurrent Validity (Pearson Correlation)

	AAPPL Reading and Character Recognition	AAPPL Reading and Word Recognition	AAPPL Reading and Reading Comprehension
Grade 3	.490**	.499*	.449**
Grade 4	.665**	.683**	.544**
Grade 5	.650**	.633**	.630**

Pearson correlations were calculated between AAPPL reading test scores and the MI-LMU STARTALK Character Recognition, Word Recognition and Reading Comprehension tests for each grade respectively. All correlations were statistically significant ($p < .01$, ranges from .49 to .68), indicating that there is a strong correlation between our assessments and AAPPL reading tests, while noting that our correlation indices are not very high, possibly because of the fact that our tests mainly focus on vocabulary and character knowledge while the AAPPL interpretive reading test focuses on reading proficiency, which are two separate constructs.

As result of these analyses, it can be determined that our assessments have strong criterion validity. As an added level of confirmation of validity, we engaged an independent analysis of the overall alignment of our suite of assessments with ACTFL interpretive reading proficiency as measured by the AAPPL Interpretive Reading results of our study (Appendix III). Prepared by two ACTFL experts, the report concludes that of the 325 grades 3-5 students, a significant number of students (N=264; 81%) are performing at or above novice-high levels of reading comprehension. These results support the validation results we found, particularly in the notion that character and vocabulary recognition have a strong correlation with global reading proficiency in Chinese.

Phase III- Grades 3-5 Reading Comprehension Assessments

Participating Schools, Grades and students

A total of 12 teachers from nine schools participated in the field test, which included four Grade 3 classes, four Grade 4 classes and four Grade 5 classes. A total number of 543 students registered for our assessments. Participating schools comprised public, charter, and independent schools from California, Utah, Minnesota, and Washington DC. Below is a synopsis of the registered number of students by school and grade.

Number of Registered Students by Grade and School

	Grade 3	Grade 4	Grade 5
Broadway	83	80	
CAIS		33	16
DLS			47
Lone Peak		54	
North Park	45		
Stewart			54
Yinghua	25	26	
Yuming			
Yuying	36		44
Total	**189**	**199**	**161**

Data Analysis and Results for Phase III Comprehension Assessments

All assessment data were collected from participating teachers between November 2017 and January 2018. Item-level scores of AAPPL reading test results were gained from ACTFL. These data were then cleaned and categorized based on grade level and test type before being evaluated through reliability and validity testing.

Given that some students who registered were not able to take the assessments due to illness or absence, invalid cases were removed from the original data sets by grade and test type to avoid missing data. The following table presents a final summary of students who completed

each test, which were submitted for reliability and validity analysis. In addition, 405 students took the AAPPL interpretive reading test, whose results were then used for concurrent validity testing.

Descriptive Statistics: Number of students completed each test

Grade	School	Language Arts	Social Science	Math
3	Broadway	81	80	81
3	Yuying	36	36	36
3	Yinghua	25	25	25
3	NorthPark	45	45	45
4	Broadway	78	58	80
4	CAIS	15	16	16
4	Lone Peak	54	54	54
4	Yinghua	26	26	26
5	CAIS	16	16	15
5	Denver Language School	47	47	47
5	Yuming	37	35	38
5	Stewart	54	54	51
Total		514	492	514

Reliability Results

Reliability is the notion that the test produces stable and consistent results over time. For the current project, reliability tests were conducted based on question format. Specifically, Rasch analysis was performed to decide item reliability of multiple choice and fill-in-the-blank questions, whose answers are dichotomous. Rasch analysis has been applied to assessments in a wide range of disciplines. The Rasch model is the only item response theory model in which the total score across items characterizes a person totally. By conducting Rasch analysis, the

total item reliability for each test as well as the fitness/unfitness of specific items in the test were obtained. In this way, poor-quality test items could also be diagnosed. Six Rasch analysis were run for corresponding tests and grades, the results of which can be found in the following table.

Rasch Analysis

Grade	Multiple Choice Items	Fill-in-the-blank Items
3	Item reliability 0.95	Item reliability 0.97
4	Item reliability 0.94	Item reliability 0.97
5	Item reliability 0.96	Item reliability 0.89

According to the above results, all of the Multiple Choice and Fill-in-the-blank item reliability scores were above 0.94, which shows directly that these test items are of very high quality. Specifically, these data demonstrate that the difficulty levels of the items are variant enough to differentiate participants' abilities. The only exception is Grade 5 Fill-in-the-blank test, whose item reliability score is relatively lower at 0.89. However, this figure is still close to the threshold of 0.90. The major reason is that only 12 fill-in-the-blank items were developed for Grade 5 test, which is difficult to reflect the variance of the items. Considering the limited number of items, the result can be considered satisfactory.

Coefficient alpha (i.e., Cronbach's alpha), which is the most popular way to estimate test reliability, was calculated for short answer questions. It measures the extent to which the items provide consistent information regarding the students' mastery of the domain. Coefficient alpha was calculated based on the number of items on the test, proportion of examinees who answered each item correctly, and sample variance for the total score. Results were summarized in the following table.

Short Answers Items: Cronbach's Alpha

	Grade 3	Grade 4	Grade 5
Short Answer Reliability	0.944	0.970	0.978

As can be seen from the table, all the test reliability indexes are well above .90, indicating high consistency of the test items.

Validity Results

We used content validity, concurrent validity as well as construct validity in order to determine the degree to which there is a match between test questions and what they are intended to assess. Content validity answers the question: Do our assessments measure reading comprehension of the intended content areas (i.e., language arts, social science, and math) for Grade 3-5 Chinese language learners? Content validity was achieved with Chinese literacy content experts in the following ways:

1. All test items were developed by master teachers who are domain experts in the content;
2. Language and literacy domain experts provided training to these master teachers, ensuring that items were selected from the appropriate content;
3. All items were cross-examined by peer master teachers from different schools.

Concurrent validity was measured to assess whether our assessments have strong criterion validity--namely whether our assessments measure Chinese reading proficiency. The AAPPL Reading test was utilized as the benchmark test with which to establish concurrent validity. As a result, correlations between AAPPL test scores and scores in our tests were calculated in order to measure concurrent validity. It is important to note that the AAPPL Interpretive reading test items are not aligned to content standards nor are they grade-level specific. Also, AAPPL interpretative reading test items do not include discrete items as the test items we developed here. Thus, there is as expected variance in our results in using it as our high stakes test. The concurrent validation results using the AAPPL are presented in the following table.

Concurrent Validity (Pearson Correlation)

	AAPPL Reading & MLA	AAPPL Reading & Social Science	AAPPL Reading & Math
3	0.578***	.439***	.414***
4	.447***	.356***	.406***
5	.754***	.743***	.670***

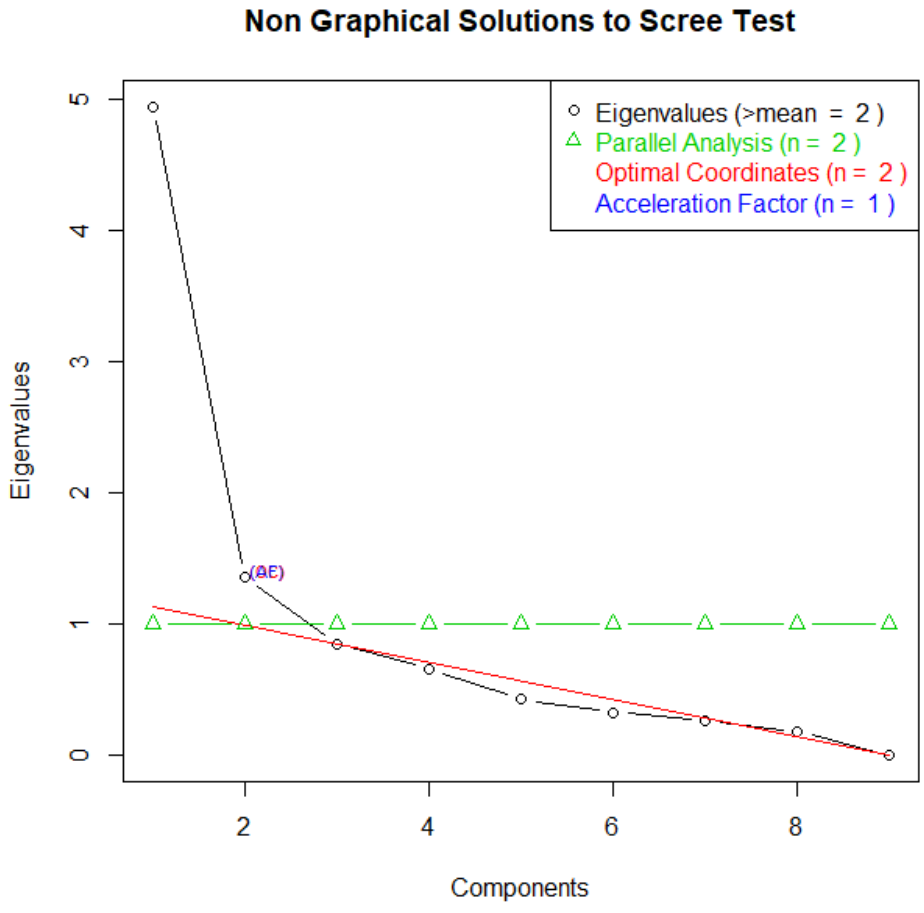
Pearson correlations were calculated between AAPPL reading test scores and the Mandarin Institute -LMU STARTALK Reading Comprehension tests by content areas for each grade respectively. All correlations were statistically significant ($p < .01$, ranges from .36 to .75), indicating that there is a strong correlation between our assessments and AAPPL reading tests. Grade 5 items are of strongest correlation with AAPPL, indicating a relatively high criterion validity.

Construct validity assess whether the developed items are testing the intended construct (i.e., Grade 3-5 Chinese learners' reading proficiency in this context). In order to answer this question, nine components of each grade's test were extracted based on their content area and question format, which are language arts multiple choice items, social science multiple choice items, math multiple choice items, language arts fill-in-the-blank items, social science fill-in-the-blank items, math fill-in-the-blank items, language arts short answer items, social science short answer items, and math short answer items. A correlation matrix of these 9 variables was created, and the analysis of which would reveal the number of latent constructs of these test items. The hypothesis is that if the test items have construct validity, their variances should only be explained by one underlying construct – reading proficiency, or by content areas. However, if the variance is explained by the test formats, it indicates that there is lack of construct validity.
A Factor analysis was run separately for each grade and the results are shown below.

Grade 3

	MLA MC	SC MC	Math MC	MLA Blank	SC Blank	Math Blank	MLA QA	SC QA	Math QA
Loadings on PA1	0.78	0.77	0.58	0.69	0.81	0.6	0.63	0.81	0.62

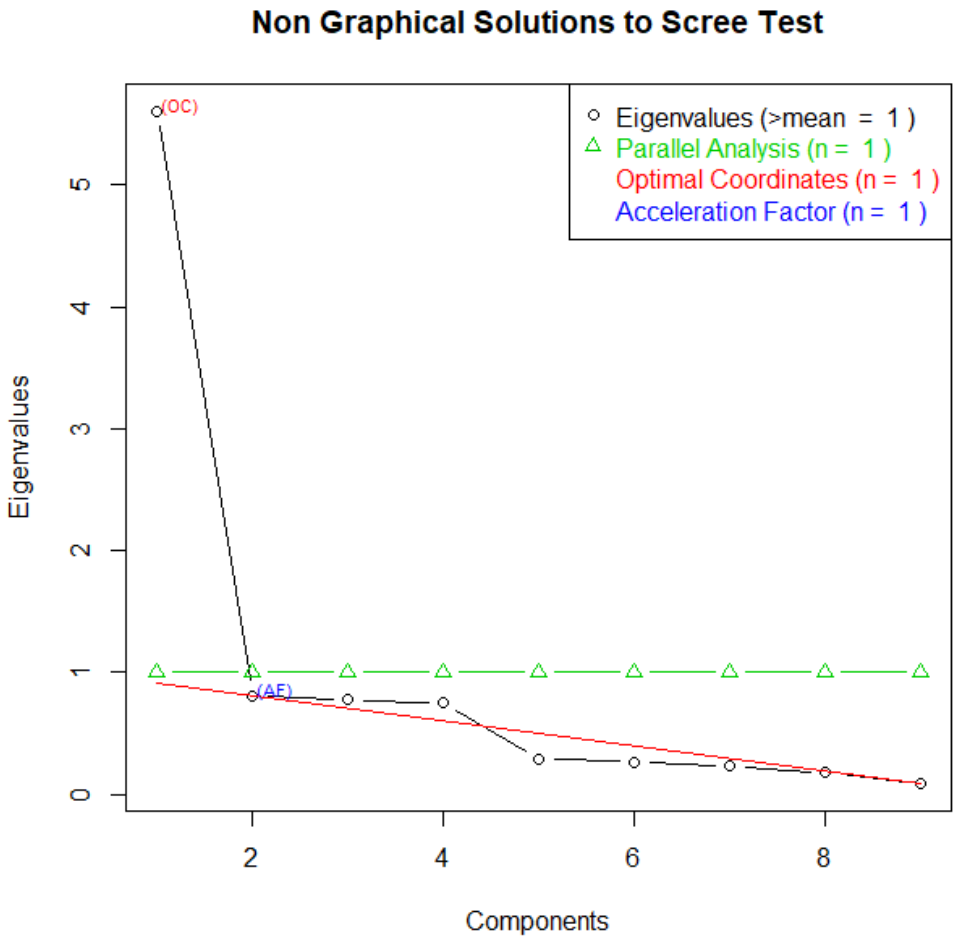
SS Loadings	4.46
Proportion Variance	0.50
Correlation of scores with factors	0.95
Multiple R square of scores with factors	0.91
Minimum correlation of possible factor scores	0.82



Grade 4

	MLA MC	SC MC	Math MC	MLA Blank	SC Blank	Math Blank	MLA QA	SC QA	Math QA
Loadings on PA1	0.85	0.85	0.63	0.48	0.86	0.46	0.88	0.92	0.78

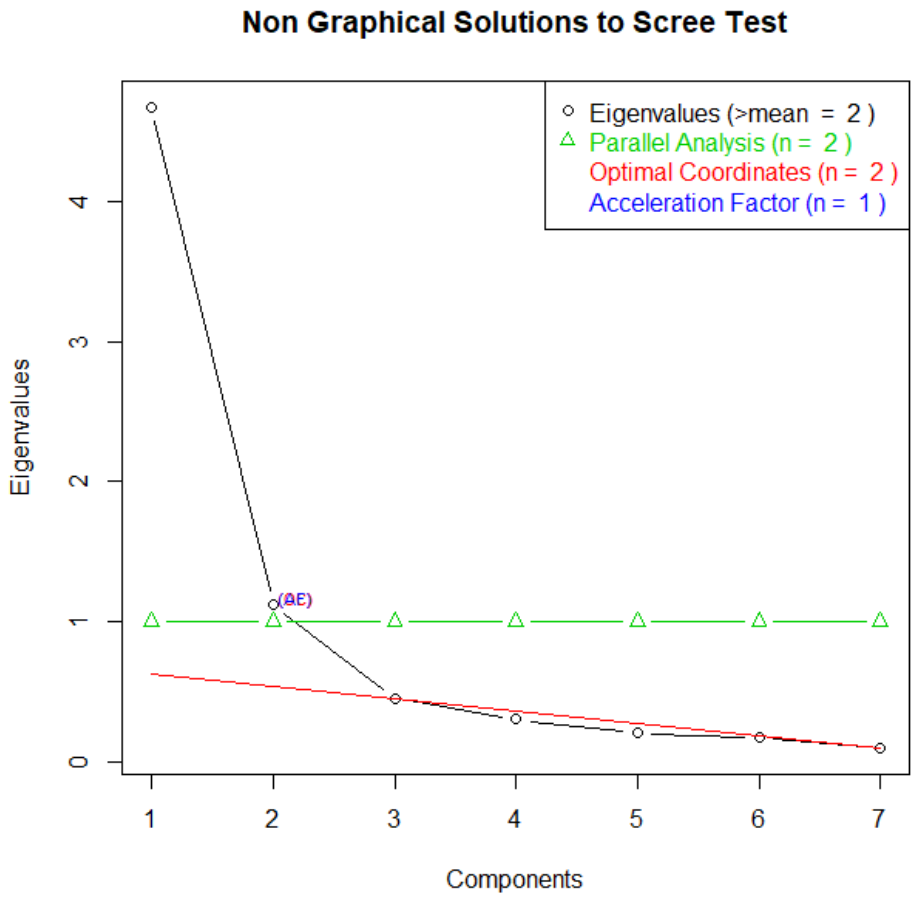
SS Loadings	5.27
Proportion Variance	0.59
Correlation of scores with factors	0.97
Multiple R square of scores with factors	0.95
Minimum correlation of possible factor scores	0.90



Grade 5

	MLA MC	SC MC	Math MC	Math Blank	MLA QA	SC QA	Math QA
Loadings on PA1	0.77	0.73	0.28	0.33	0.91	0.88	0.25
Loadings on PA1	0.38	0.38	0.69	0.87	0.22	0.33	0.82

	PA1	PA2
SS Loadings	2.43	2.35
Proportion Variance	0.43	0.34
Correlation of scores with factors	0.96	0.94
Multiple R square of scores with factors	0.92	0.88
Minimum correlation of possible factor scores	0.85	0.76



Principal Axis Factor Analysis reveals that there is one underlying factor explaining all 9 variables in Grade 3 and Grade 4 test items. In other words, that is one construct "reading proficiency", implying that the test items have construct validity, given that their variances are not accounted for by either content area or method, but all by learners' proficiency.

The result is a bit different in Grade 5 test. Principal Axis Factor Analysis reveals that there are two underlying factors. While all Chinese and Social Science items load onto the first factor, Math items load onto the second factor. In other words, two underlying constructs explain the variance of all the components: Chinese/Social Science and Math. The correlation between Chinese and Social Science is quite high, which can be categorized as Reading, while in this case, Math items seem to assess different skills. Nevertheless, the results are still considered very strong evidence for the construct validity the assessments, since the variability is accounted for by content area but not by method of testing.

Conclusions

In summary, by the end of Phase III, field testing data reveal that the character and vocabulary recognition and reading comprehension assessments developed at the STARTALK Institute demonstrate high reliability and validity. Since test items were selected based on the K-5 Chinese Word Frequency Dictionary for L2 Learners, these results also provide evidence that both the WFD as well as the assessments are quite reliable and valid.

One of the limitations of Phase II was that since greater focus was given to character and vocabulary tests, this resulted in a limited number of items in reading comprehension tests. Reliability results were very strong; concurrent validation results, as revealed in the AAPPL correlation indices with the AAPPL test were relatively lower, yet still significant.

Our work continued during Phase III, when the focus on assessing reading ability by expanding the number of items for reading comprehension as well as test types through a rigorous statistical analysis process that will strengthen the field of early Chinese literacy and assessments. The results of field testing of data revealed that all reading comprehension assessments developed at the STARTALK Institute demonstrate high reliability and validity. This collection of tests not only covers a broad range of content areas, but also include different test

formats. The evidence of criterion validity and construct validity also implies that these tests would be very informative tools of assessing learners' reading proficiency.

One of the limitations of the current project is that while a balanced number of test items in each format was intended at first, it turned out that there was a lack of items in fill-in-the-blank type. The assessments would be more comprehensive if more fill-in-the-blank items are included.

In addition, one of the challenges for this work is to compare them with any existing standard tests. Any generalization or conclusion based on these assessments should be made with caution, however, by collaborating with proficiency and literacy experts, and in-service master practitioners, we are able to develop instruments that will contribute to the field at large.

The development of the K-5 Chinese Word Frequency Dictionary for L2 Learners and this set of assessments developed across the three Phases of STARTALK Institutes and funding are the result of pioneering work that was enabled by STARTALK. The character, vocabulary and reading comprehension assessments developed and validated are technically sound and can serve as baseline instruments and tools to determine Chinese early literacy across a variety of immersion program types. This initial set of assessment and corresponding validation reports were the result of multiple years of collaboration between language proficiency and literacy experts, and in-service master practitioners, we are able to develop instruments that will contribute to the field at large.