Avant STAMP™ 4S (STAndards-based Measurement of Proficiency – 4 Skills)*

Chinese Technical Report

By Martyn Clark, Assessment Director

Center for Applied Second Language Studies (CASLS)

Updated by Dr. Jim Snyder, Director of Market Research

Avant Assessment LLC

7/16/2012

**Avant STAMP 4S is a proficiency-oriented assessment of listening, reading, writing and speaking*

NOTICE: The contents of this report were developed under a grant from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

This report is an updated version of a previous report titled *Chinese Computerized Assessment of Proficiency (Chinese Avant STAMP 4S)* published by CASLS (Technical Report 2009-1). The writing and listening scoring section was updated to reflect Avant Assessment's process and the previous report included some additional test functionality that was not included in the current Avant STAMP 4S delivered by Avant Assessment.

Abstract

This document was prepared by the Center for Applied Second Language Studies (CASLS) and updated by Avant Assessment. It describes the development of Avant STAMP 4S in Chinese. The development of the test was made possible by the University of Oregon Chinese Flagship with funding from the National Security Education Program (NSEP). Some additional funding was provided the Department of Education through the Title VI program. Avant STAMP 4S is an online proficiency-oriented test of listening, reading, writing and speaking.

This document has six major sections. The first is an overview of the Chinese Avant STAMP 4S project. The second section describes the assessment. The third section details the development of the test items. The fourth describes the test's technical aspects. The fifth section discusses validity evidence associated with the test. The final section presents information about appropriately interpreting scores from the test.

Acknowledgment

The development of this test was made possible by the University of Oregon Chinese Flagship with funding from the National Security Education Program (NSEP). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSEP. Additional materials were developed under a grant from the U.S. Department of Education. However, those contents do not necessarily represent the policy of the U.S. Department of Education, and you should not assume endorsement by the Federal Government.

Contents

Nomenclature

ACTFL	American Council on the Teaching of Foreign Languages
Avant	Avant Assessment
Bin	A group of test items delivered together
CASLS	Center for Applied Second Language Studies
FSI/ILR	Foreign Service Institute/Interagency Language Roundtable
Item set	Two or more items sharing a common stimulus (e.g., a reading text)
LRC	Language Resource Center
Level	Level on a proficiency scale (e.g., Advanced-Mid)
Panel	A term used to describe a particular arrangement of bins
Rasch	A mathematical model of the probability of a correct response which takes person ability and item difficulty into account
Routing table	A lookup table used by the test engine to choose the next most appropriate bin for a student
Score table	A lookup table used by the scoring engine to determine an examinee's score based on their test path
STAMP 4S	STAndards-based Measurement of Proficiency - 4 Skills
Test path	A record of the particular items that an examinee encounters during the test

Preface

The Center for Applied Second Language Studies (CASLS) is a Title VI K-16 National Foreign Language Resource Center at the University of Oregon. CASLS supports foreign language educators so they can best serve their students. The center's work integrates technology and research with curriculum, assessment, professional development, and program development.

CASLS receives its support almost exclusively from grants from private foundations and the federal government. Reliance on receiving competitive grants keeps CASLS on the cutting edge of educational reform and developments in the second language field. CASLS adheres to a grass-roots philosophy based on the following principles:

- All children have the ability to learn a second language and should be provided with that opportunity.
- Meaningful communication is the purpose of language learning.
- Teachers are the solution to improving student outcomes.

Avant STAMP 4S is an online test of proficiency developed by CASLS. In the past, proficiency tests developed at CASLS have been licensed by Avant Assessment through a technology transfer agreement overseen by the University of Oregon Office of Technology Transfer. These tests are delivered operationally under the name Avant STAMP 4S (STAndards-based Measurement of Proficiency – 4 Skills).

Avant Assessment LLC, founded in 2001, set out to become a world leader of innovative language assessment solutions by merging expertise in assessment, linguistics and technology. Avant's founders and current leadership believe in a world where language is no longer a barrier, a world where every teacher is able to describe with confidence the strengths and needs of every student in their care, and a world where every student has accurate evidence of their educational abilities and can set goals to match their needs and interests. Our commitment is to provide meaningful data and evidence that inspires that confidence.

Executive Summary

CASLS has developed Avant STAMP 4S in Chinese, an online assessment of Mandarin Chinese that covers a proficiency range comparable to the American Council on the Teaching of Foreign Languages (ACTFL) proficiency levels Novice through Advanced in 4 skills (listening, reading, writing and presentational speaking). This test builds on the style and format of the Chinese Standards-based Measurement of Proficiency (STAMP) test created previously at CASLS. The Avant STAMP 4S project introduces a new item development process, an additional skill, and a new delivery algorithm for the listening and reading sections.

Native Chinese speakers identified or constructed listening and reading passages and CASLS staff wrote test items according to task specifications. A comprehensive external review of the test items was conducted in August 2008. Reviewers expressed general satisfaction with the test items, and there was a high correlation ($r = .87$) between the intended proficiency target level of the items and the expert reviewers' ratings.

The best reviewed items were arranged into a test panel for pilot testing. More than 1,000 learners in programs across the country participated in pilot testing. Analysis of the pilot data showed reliabilities of .95 and .92 for the listening and reading section, respectively. Cut scores for the major proficiency levels were determined, and a subset of piloted items was arranged into bins for operational multistage adaptive delivery. Simulation studies of the delivery algorithm show a correlation of $r = .97$ between simulated test taker ability and final ability estimate on the operational version of the test.

1 Overview and purpose of the assessment

1.1 Construct for Avant STAMP 4S

Avant STAMP 4S can be considered primarily a "proficiency-oriented" test. Language proficiency is a measure of a person's ability to use a given language to convey and comprehend meaningful content in realistic situations. Avant STAMP 4S is intended to gauge a student's linguistic capacity for successfully performing language use tasks. Avant STAMP 4S uses test-taker performance on language tasks in different modalities (speaking, reading, listening and writing) as evidence for this capacity.

In Avant STAMP 4S, genuine materials and realistic language-use situations provide the inspiration for the listening and reading tasks. In many cases, authentic materials are adapted for the purposes of the test. In other cases, these materials provide the template or model for materials created specifically for the test. Listening and reading items are not developed to test a particular grammar point or vocabulary item. Rather, the tasks approximate the actions and contexts of the real world to make informal inferences as to how the learner would perform in the "real world."

1.2 Test level

CASLS reports assessment results on the CASLS Benchmark Scale. Several points along the scale have been designated as Benchmark Levels. These Benchmark Levels include verbal descriptions of the proficiency profile of a typical student at that point in the scale.

The Benchmark Level descriptions are intended to be comparable to well-known proficiency scales at the major proficiency levels, notably the FSI/ILR scale and the ACTFL Proficiency Guidelines, as these are used widely. The conceptual relationship between the scales is shown in Table 1, with sub-levels shown for completeness. Correlations between CASLS' intended proficiency levels and levels based on expert review can be found in Section 5.3 on page 20.

The following verbal descriptions characterize proficiency at each of the CASLS Benchmark Levels:

Level 3 (Beginning proficiency) Beginning proficiency is characterized by a reliance on a limited repertoire of learned phrases and basic vocabulary. A student at this level is able recognize the purpose of basic texts, such as menus, tickets and short notes, by understanding common words and expressions. The student is able to understand a core of simple, formulaic utterances in both reading and listening. In writing and speaking, the student is able to communicate basic information through lists of words and some memorized patterns.

Table 1
CASLS Benchmark Levels

Benchmark	CASLS Level	ILR	ACTFL
Refining	Level 10	3	Superior
Expanding	Level 9	2+	Advanced-High
	Level 8		Advanced-Mid
	Level 7	2	Advanced-Low
Transitioning	Level 6	1+	Intermediate-High
	Level 5		Intermediate-Mid
	Level 4	1	Intermediate-Low
Beginning	Level 3	0+	Novice-High
	Level 2		Novice-Mid
	Level 1	0	Novice-Low

Level 5 (Transitioning proficiency) Transitioning proficiency is characterized by the ability to use language knowledge to understand information in everyday materials. The learner is transitioning from memorized words and phrases to original production, albeit still rather limited. In reading, students at this level should be able to understand the main ideas and explicit details in everyday materials, such as short letters, menus, and advertisements. In listening, students at this level can follow short conversations and announcements on common topics and answer questions about the main idea and explicitly stated details. In speaking and writing, students are not limited to formulaic phrases, but can express factual information by manipulating grammatical structures.

Level 8 (Expanding proficiency) Expanding proficiency is characterized by the ability to understand and use language for straightforward informational purposes. At this level, students can understand the content of most factual, non-specialized materials intended for a general audience, such as newspaper articles and television programs. In writing and speaking, students have sufficient control over language to successfully express a wide range of relationships, such as temporal, sequential, cause and effect, etc.

Level 10 (Refining proficiency) Refining proficiency is characterized by the ability to understand and use language that serves a rhetorical purpose and involves reading or listening between the lines. Students at this level can follow spoken and written opinions and arguments, such as those found in newspaper editorials. The students have sufficient mastery of the language to shape their production, both written and spoken, for particular audiences and purposes and to clearly defend or justify a particular point of view.

The four Benchmark Level labels can be remembered by the mnemonic BETTER (BEginning, Transitioning, Expanding and Refining).

Chinese Avant STAMP 4S currently measures students up through the Expanding Level (ACTFL Advanced / ILR Level 2). Table 2 shows a detailed description of the language construct for Chinese Avant STAMP 4S.

Table 2

Language Proficiency Measured by Avant STAMP 4S (based on Bachman & Palmer (1996))

		Beginning	Transitioning	Expanding	Refining
Grammar	Vocabulary	knowledge of limited number of common words and cognates	knowledge of some general purpose vocabulary	knowledge of most general purpose vocabulary and common cultural references	knowledge of general purpose vocabulary and some specialized vocabulary
	Syntax	little productive ability, but may be able to recognize memorized chunks	familiarity with basic syntactic structures, but no complete accuracy; may be confused with complex structures	familiarity with basic syntactic structures and common complex constructions	generally able to understand all but the most complex or rare syntactic structures
Text	Cohesion	little or no cohesion	some knowledge of cohesion, but may be confused by relationships	able to recognize and express most common relationships (temporal, sequential, cause and effect, etc.)	able to understand a wide range of cohesive devices
	Rhetorical Organization	loose or no structure	loose or clear structure	able to recognize clear, underlying structure	able to recognize structure of argument
Pragmatic	Functional	ability to recognize basic manipulative functions	ability to understand basic manipulative and descriptive functions	heuristic (language for learning)	imaginative (language used to create imaginary worlds, poetry)
	Sociolinguistic	combination of natural and contrived language	combination of natural and contrived language	mainly natural language	able to recognize register differences, figures of speech, etc.

Note: Topical knowledge and Strategic knowledge are not explicitly assessed, but test takers are expected to have general knowledge of the world and some test takers may be able to make use of test-taking skills

1.3 Population served by the assessment

Description of the test taker

The target audience for this test is adult (age 13+) language learners. The test takers are assumed to be native English speakers or to have a high degree of fluency in English and to be literate. The test takers will be primarily students in programs that teach Modern Mandarin Chinese, but they may also be persons seeking to enter such programs, including those who have learned the language informally.

Description of the test score user

Examinees, language instructors and program administrators are the intended score users. Examinees will use the test score to evaluate their progress toward their language learning goals. Language instructors will use the scores, in conjunction with multiple other sources of information, to help inform placement decisions and evaluations. At the class level, aggregate information can help inform curricular decisions for program administrators.

Intended consequences of test score use

The ultimate goal of the test is to increase the foreign language capacity of language learners in the U.S. As such, it is hoped that use of the test positively influences programs in terms of putting a greater value on proficiency and meaningful language use, as opposed to rote memorization.

CASLS and Avant Assessment suggest that educators not use Chinese Avant STAMP 4S (or any other single assessment) as the sole basis of making decisions affecting students. These decisions might include graduation and credit issues. Used in connection with other measures, such as course grades, teacher evaluations and other external assessments, Avant STAMP 4S can help provide additional empirical data on which to base decisions.

2 Description of the assessment

Chinese Avant STAMP 4S is designed to provide a general overall estimate of a language learner's proficiency in four skills in modern Mandarin Chinese. The test is delivered via the Internet without the need for any special software. It is a snapshot of language ability based on a relatively short number of tasks. As such, the Avant STAMP 4S is not a substitute for the judgment of an experienced classroom teacher, nor is it sensitive enough to make high-stakes claims regarding a test taker's language proficiency. Avant STAMP 4S can be used effectively, however, to gauge general proficiency at the start of a course to inform placement decisions or to provide an indication of general proficiency at the end of a course for summative assessment. Because Avant STAMP 4S results are reported on a scale consistent with the widely used ACTFL and ILR proficiency scales, it can provide a common touchstone for comparison at the school, district, or state level. A foreign language instructor knows his or her students the best, but does not necessarily know how those students compare to students in similar programs in other places; a standardized assessment like Avant STAMP 4S can help facilitate such comparisons.

2.1 Content and structure of the Avant STAMP 4S

The Chinese Avant STAMP 4S consists of four sections:

- Interpretive Listening
- Interpretive Reading
- Presentational Writing
- Presentational Speaking

The listening and reading sections consist of multiple-choice items and are scored automatically by the test engine. In the writing and speaking sections, examinee performance data is captured by the computer and saved to a database where a trained external rater from Avant Assessment rates the work according to a simple rubric (See Section 4). Although the different sections of Avant STAMP 4S are meant to work together to give a snapshot of the examinee's overall proficiency, the sections themselves are scored separately and can be delivered in a modular fashion. There is no aggregate score on Avant STAMP 4S. This is done to give language programs the maximum flexibility in using the test. Programs can choose to use all sections of Avant STAMP 4S outright or can choose specific sections to supplement assessment practices already in place.

A typical item on the Chinese Avant STAMP 4S reading test may look something like Figure 1. Examinees are presented with a situation that describes a realistic language use context. A graphic contains both the Chinese text as well as contextualizing information. The test question, in English, requires the examinee to read the information in Chinese and choose the best answer from the options provided. Examinees must answer the question before proceeding to the next screen. Backtracking is not allowed.

Figure 1. Chinese reading item

Chinese listening items (Figure 2) are similar to their reading counterparts. Examinees are presented with a situation in English that describes a realistic language use context. The audio playback button allows examinees to start the audio stimulus when they are ready. Once the audio begins playing, it will play until the end of the file. Once the playback button has been pressed twice it will no longer be active. Examinees can hear the audio only twice per item. As with the reading section, backtracking is not allowed and examinees must answer the question before proceeding. If a particular audio passage has more than one associated item, examinees will be able to play the audio twice for each of the associated items if they choose.

2.2 Test delivery

The Chinese Avant STAMP 4S is delivered over the Internet using any standard browser. The login scheme is based on classes, and it is assumed that most students taking the test will do so in a proctored environment, such as a computer lab. The listening and reading sections of Chinese Avant STAMP 4S is delivered using a multistage adaptive testing paradigm (Luecht, Brumfield, & Breithaupt, 2006; Luecht, 2003). Items in the test are arranged into multi-item testlets or bins of different difficulties. As the examinee completes one bin of items, the next bin is chosen based on how well he or she performed on the previous bin. Examinees who got most of the items correct will receive more challenging items in the next bin, while examinees who did not do so well will receive items at the same level.

Figure 2. Chinese listening item

For operational Chinese Avant STAMP 4S delivery, a multistage delivery is used (Figure 3). Because one of the primary funded goals of the Chinese Avant STAMP 4S was for use in Flagship programs, this algorithm provides the most efficient method of delivering the test items. There are several trial items embedded in the operational test, but these do not count towards the final score. The particular delivery configuration used to pilot Chinese Avant STAMP 4S has been termed the "Floor First" model because it required examinees to do well on easy items before being challenged with more difficult ones (see Figure A.1 in Appendix A).

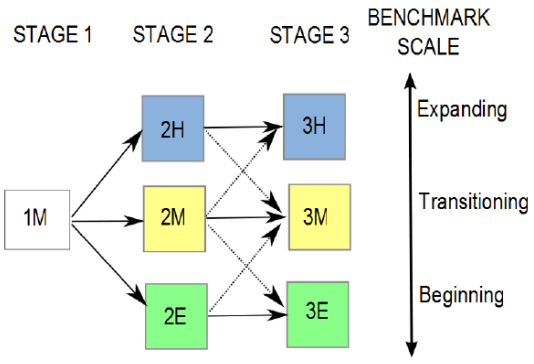

Figure 3. Multistage adaptive algorithm

3 Test development

The content for Chinese Avant STAMP 4S was created in two separate phases. Items covering the Beginning and Transitioning level (ACTFL Novice and Intermediate) were developed by CASLS staff and partners between 2002 and 2006 as part of a Language Resource Center (LRC) grant. A sample of the CASLS Benchmarks upon which these original items were developed is presented in Figure B.2 in Appendix B. CASLS allocated additional funding to develop the Expanding and Refining levels of the test between 2007 and 2009. This development coincided with a reworking of the entire assessment framework, including test design and delivery[1]. The development process for this most recent phase of the test is illustrated in Figure 4. Major components of this process are described below.

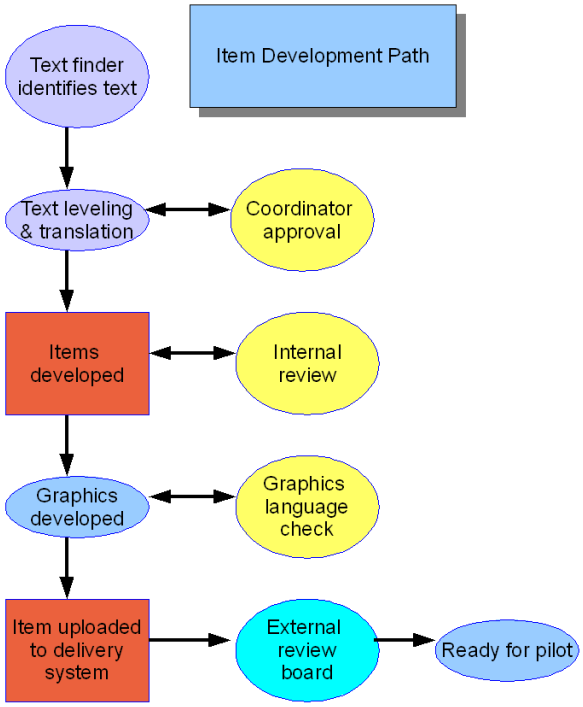

Figure 4. Item writing workflow

3.1 Text finding

CASLS hired two Chinese-speaking graduate students as text finders to work alongside Chinese speaking CASLS staff to find reading texts and produce draft items for this project. These text finders were given training on rating passages according to ILR levels via a CD-ROM-based passage rating course produced by the National Foreign Language Center (NFLC). The passage rating system is based on the work of Child (1987, 1998) and describes the features of texts thought to exemplify four increasingly complex modes of written communication. Of the texts collected by the text finders, only a subset were suitable for item development. In particular, listening texts appropriate for the Refining (ACTFL Superior) level

[1] Detailed test and task specifications are available on the CASLS website at http://casls.uoregon.edu.

proved difficult to find. As a result, the bulk of the items were developed for the Expanding (ACTFL Advanced) level.

3.2 Internal review

Approximately halfway through the project, CASLS hired a full-time Test Developer. Chinese items were reviewed by the CASLS Assessment Director and Test Developer (working from English translations), and feedback was given to the text finders. At this stage of the process, some passages and items were determined to be inappropriate for the test because of content or required background knowledge and were not developed further. In addition, as CASLS clarified and updated the task specifications, the Test Developer took over more responsibility for item writing, working closely with the text finders to ensure that the resulting items were appropriate for the passages and intended proficiency levels.

3.3 Graphics development

Because the test is intended to be compatible with any computer, CASLS renders the Chinese text as a graphic to avoid any font display issues when the test is delivered (see sample item on page 11). For each text on the test, CASLS graphic artists imported a screenshot of the original word processor text into context-appropriate images that were then uploaded to the test delivery system. The Chinese-speaking text finders reviewed finished items to ensure that the text was being correctly displayed in the final item.

Table 3 on page 23 shows the number of items produced for the project, including those developed through previous work on the test.

3.4 Revisions

After the external review (see Section 5), CASLS staff selected the most promising texts and items to appear on a pilot version of the test. This opportunity was also used to group previously created items related to the same text into item sets[2]. To avoid dependencies across these items, only those questions in each set considered independent were activated for test delivery.

Table 3

Item Counts for Chinese

Level	Simplified Reading	Traditional Reading	Listening
Beginning	69	47	111
Transitioning	85	64	89
Expanding	51	51	34

[2] Set-based delivery was not available in early incarnations of the test engine when some of the items were developed.

4 Technical characteristics

4.1 Selection of items

Not all items developed for the test have been included in the operational form. Responses from a pilot testing questionnaire (see Appendix G), and comments from test users indicated that the length of the pilot test (between 30 and 60 items, depending on the student) was too long to be a workable solution in many situations. Much of this variable test length is a function of the "Floor First" algorithm that was used for piloting (see Appendix A). Even though shortening the test would decrease the precision of the scores, it was decided that a shorter version would be much more usable by the target user group.

Data from pilot testing was analyzed using the Rasch model (Rasch, 1960/1980) as implemented in Winsteps (Linacre, 2008). In the test delivery system, Chinese is treated as two separate languages, depending on whether the simplified character set or traditional character set is used for the reading passages. Because educators assume that scores from the two versions will be identical, it was decided that combining the data for the analysis and calibration would be important. Before combining the data, each test was analyzed separately and the item difficulties crossplotted (see Appendix C for an example). Seven reading items showed discrepancies between the traditional and simplified version of the items, and those were removed from the combined analysis[3]. Four listening items were also removed from the combined analysis based on an analysis of the listening crossplot.

Summary results of the combined analyses are shown in Appendix D. The Rasch person reliabilities of the listening and reading sections were .95 and .92 respectively, and most items showed good fit to the model. Items with mean squared infit values between .5 and 1.5 were considered acceptable for inclusion in the pool. The difficulty values for these items will be used as anchor values when calibrating new items into the pool in the future.

4.2 Preparation for delivery

An iterative process was used to place items in bins for multistage delivery. The goal was to create bins of 10 items each. However, because many items were part of an item set, it was not always possible to create the optimum arrangement that would maximize the information[4] for each bin (see Figure 4). For this reason, it was not possible to keep the final bin size to 10 across all of the bins. Once bins were finalized, routing tables and score tables were produced with Winsteps by anchoring item difficulties at their calibrated values and using dummy student records. The routing table is a lookup table that shows an estimated Rasch ability score for every possible raw score for every possible path through the bins. As the test progresses, examinees are routed to the most maximally informative bin (Figure 4) for their particular estimated ability at that point in the test.

[3] The correlation between traditional and simplified reading item measures was r = .94, disattenuated r = .98.

[4] The information function for a bin is the sum of the individual item information functions.

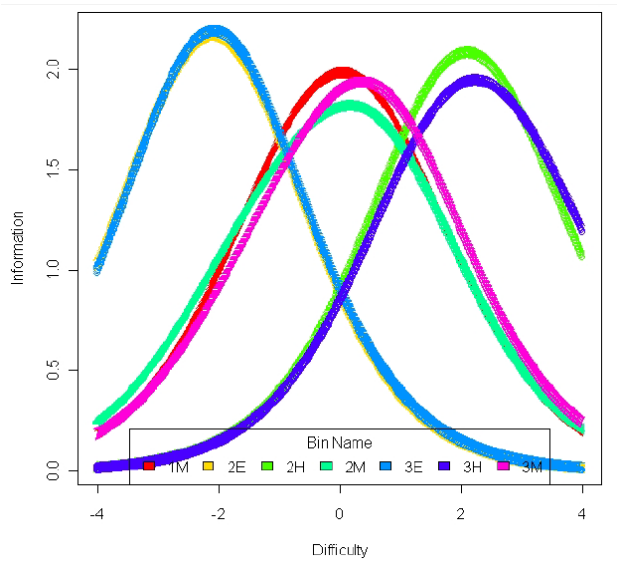

Figure 4. Information functions for Chinese listening bins

Similarly, the score table is a lookup table that gives a final estimated ability for every possible raw score for every possible test path.

4.3 Determination of cut scores

Cut scores for the Chinese Avant STAMP 4S were set using an 80% success criteria. It is important to set the cut scores relative to the Rasch ability continuum rather than relative to any particular set of items. In this way, future versions of the test can maintain cut scores that are consistent across time. To set the cut scores, the median difficulty was calculated for items in each of the three proficiency groups, Beginning, Transitioning, and Expanding[5]. This difficulty level represents the point at which an examinee has a 50% chance of getting an item of median difficulty correct[6]. To represent an 80% probability of success, 1.4 logits was added to the median value for each level to produce the final cut score. Examinees that have an ability estimate equal to the cut score will have an 80% or probability of success on median difficulty items for that level. Note that all of the items used in the setting of cut scores were items that the external reviewers had previously identified as being appropriate for the targeted proficiency level.

The cutscores for the test can be found in Table 5 in Section 6.2.

4.4 Test simulations

A simulation study was performed on the finalized listening and reading panels. A set of 10,000 simulated test-takers was created with abilities generated from a uniform distribution that covered the logit range of the test items. Plotting the simulated ”true” ability and the ability estimate generated from the test score table showed a strong positive relationship, with a correlation of 0.98 for listening and .97 for reading (see Appendix E).

[5] From the Rasch separation values (see Appendix D), it is possible to compute the number of strata, or statistically distinct levels of performance using the formula H = (4G+1)/3, where G is the separation index. Since neither the listening nor the reading tests had sufficient power to detect all nine proficiency levels (three main proficiency levels each with three sublevels each), cut scores were only developed for the major levels.

[6] Given by the Rasch's formula for dichotomous responses $Pr\{x_{ni}=1\} = e^{\beta_n-\delta_i}/1+e^{\beta_n-\delta_i}$, where β_n is the ability of person *n* and δ_i is the difficulty of item *i*.

To determine the simulated examinee's "true proficiency level" (in terms of the proficiency scale), a value of 1.4 was subtracted from their generated "ability" level. The resulting value's place in the range of cut scores determined the simulated examinee's "proficiency" level. The reading test was 88% on target and the listening test was 86% on target in terms of placing the student into their "real" proficiency level. This may seem low given that the test only has three possible proficiency levels (four if the undetermined level is counted), but is not unexpected as students very near the cut score for the test will be greatly influenced by the error in the test scores. For this reason, it is important to look at the scaled scores in relation to the cut scores, as well as the proficiency designation when interpreting the results of the test (see Section 6).

5 Validity evidence

A comprehensive review for the Chinese Avant STAMP 4S was held at the University of Oregon from August
10 - 12, 2008. The review aimed to:

1. have the quality of the items reviewed by independent experts, and
2. provide evidence that the items were consistent with the proficiency levels targeted by the passages.

5.1 Participants

These Chinese specialists participated as external reviewers in the review of Avant STAMP 4S:

- Dr. Jennifer Liu (University of Indiana)
- Dr. Vivian Ling (Oberlin College)
- Dr. Adam Ross (Lakeside School, Washington)
- Dr. Matthew Christensen (Brigham Young University)
- Dr. Kojo (David) Hakam (Portland Public Schools)

All of the participants were familiar with ACTFL and/or ILR Guidelines.

5.2 Procedure

The review took place over a two-and-a-half day period. The complete agenda is available in Appendix F. Day One was devoted to an overview of the test, including a review of CASLS proficiency levels and their relation to ACTFL and ILR levels. Reviewers were encouraged to ask clarifying questions about the test design, construct, and purpose.

For Day Two, a standard setting process referred to as the "Basket procedure" (Kaftandjieva, 2009) was employed. Reviewers were given full-color printouts of Chinese Avant STAMP 4S items, instructed to view the test online and, for each item, mark the minimum level of proficiency needed to correctly answer the items. (An example rating sheet is shown in Figure 5.) The order of presentation of the items was randomized by the test delivery system. To provide variety, items for both reading and listening were included in each round. Although reviewers were given the option to review the reading items using either Simplified or Traditional characters, all chose to review the Simplified version of the items. At first, each reviewer went through the items for that particular round individually, marking their estimated level on a master sheet. After each round, reviewers came together to discuss the items from that round.

The items were split into five rounds (see Table 4). Originally, it was intended that all reviewers would review every item in each round. However, after reviewing the first round it became clear that there would be insufficient time to complete the four remaining rounds as a group. To maximize the number of items reviewed, the reviewer group was split in half, with three experts reviewing Round 3 and Round 4 items and two experts reviewing Round 2 and Round 5 items. Chinese-speaking students, CASLS staff, and two representatives from Avant Assessment were present during the group discussions to take notes.

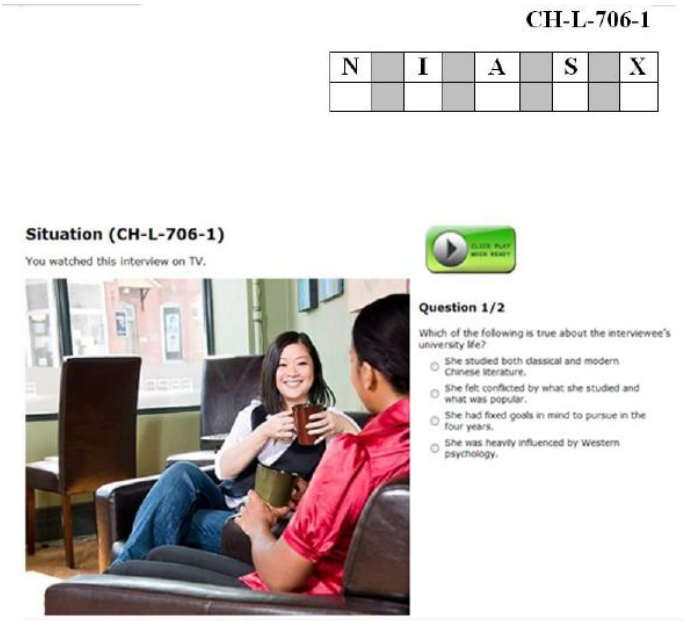

Figure 5. Review rating sheet

Table 4

Counts of Items Reviewed

Round	Reading	Listening	Total
1	26	32	58
2	27	31	58
3	40	41	81
4	43	48	91
5	43	41	84
Total	179	193	372

5.3 Results

The reviewers expressed general satisfaction with the test design and the quality of the items. The most common concern was that of mismatch between the level of the passage and level of the questions. This was most problematic at the lower levels, as the reviewers felt that beginning learners should not be taxed with "too much text on the page" even if the actual task was the recognition of a single word. An additional area of concern was the appropriateness of some of the passages for all potential test takers. The reviewers thought that some items would not be appropriate for test takers at the lower end of the age range (13+) covered by Avant STAMP 4S. Problematic items were noted for revision or exclusion.

Ratings were analyzed using multi-faceted Rasch analysis with Facets software (Linacre, 2008). This allowed the analysis of all of the items using a common frame of reference using the ratings from Round 1 to link all of the reviewers. None of the reviewers were identified as an outlier and only six standard residuals greater than 3 were observed across all of the items. The "fair average"[7] results from Facets correlated at r = 0.87 with CASLS intended item level for reading and r = 0.90 for listening.

[7] The "fair average" is the average rating on the original scale adjusted for the relative severity of the raters.

6 Score reporting

6.1 Scoring overview

Chinese Avant STAMP 4S is scored per skill. There is no aggregate score for the test as a whole. Test users should consider the information in this report when interpreting scores.

6.2 Reading and listening scores

Reading and listening scores are reported as general proficiency levels and as scaled scores. The scaled score is derived by multiplying the Rasch estimate by 45.5 and adding 500. These values were chosen to eliminate the need for decimal places in the scores. The scaled scores are simply a linear transformation of the logit scale values into a more user-friendly format and should be interpreted only in relation to cut scores for that particular skill on this test and not similar scores for other skills or other standardized tests. Cut scores for the various proficiency levels on this scaled score are shown in Table 5.

Table 5

Cut Scores for Scaled Scores

Level	Reading	Listening
Beginning	370	400
Transitioning	566	563
Expanding	665	670

There is approximately a ±21 point standard error for scaled scores. This should be kept in mind when comparing student scores or when comparing student performance to the cut scores for various proficiency levels.

6.3 Writing and speaking scores

Avant Assessment provides rating for the speaking or writing sections.

Teachers are able to log in and see their rated student items that were rated based on a simple rubric by trained Avant Assessment raters. The same rubric is used for all speaking and writing items. Writing and Speaking scores are graded by Avant-trained raters that go through a rigorous training course and are required to pass a certification test before they are allowed to rate live student responses. To insure there is Inter-Rater-Reliability, 20% of all responses are graded by a second rater and the system monitors and reports how the raters are doing with live updates of IRR. Managers monitor grading of all raters to ensure they are grading accurately and that there is no "drift" occurring. Re-training occurs on an ongoing basis and is assisted by the responses that have been flagged in the system as being scored differently by at least two raters. Avant makes every effort to ensure rating is accurate, using both computer- and human-assisted systems.

The current Avant STAMP 4S rubric is as follows:

Table 6 Avant STAMP 4S Rubric

Text Type Production	Language Control
(EB/C) – EXTENDED PARAGRAPH: Variety of cohesive devices and organizational patterns evident in response. Vocabulary is clear, specific and natural. Language is smooth and natural in delivery and without noticeable errors.	Language is fluent with limited errors. Ability to create complex language using precise and defined vocabulary. Control of the abstract as well as ease of use of idiomatic phrases and concepts. Clear, sequential ordering evident (if required) and accurately follows target-language conventions.
(EA) – PARAGRAPH: Emerging evidence of linked or connected paragraph structure. Cohesive devices used to link sentences. Complex sentence use creates depth of meaning. Increasing control of all timeframes (present, past, future, etc).	Language is error-free a majority of the time with familiar topics. If errors exist, they are patterned and do not hinder overall meaning. Delivery is mostly fluent with only occasional hesitancy. Some abstract and precise use of vocabulary and terms with familiar topics.
(TB/C) – CONNECTED: Groupings of sentences showing increased cohesion. Some use of unique and non-formulaic sentences that create deeper meaning. Use of complex sentences emerging.	Good accuracy evident with possible errors that don't affect the overall meaning. Delivery may be somewhat choppy. May have repetitive use of concrete vocabulary with occasional use of expanding terms. Accuracy for complex sentences is emerging.
(TA) – STRINGS: Able to create strings of related statements, simple questions and commands. Most formulaic sentences must have added detail (modifying phrases). Language goes beyond memorized high-frequency expressions.	Good accuracy with formulaic sentences with some added detail. Errors may occur as student attempts higher-level skills. Good control expected with majority of response.
(BC) – SIMPLE SENTENCES: Emerging ability to create simple sentences, some signs of original language emerging with errors. Often uses memorized expressions to create sentences.	Good accuracy for high-frequency expressions. Usually comprehensible to a sympathetic reader/listener. Grammatical (syntax, spelling, conjugation) errors expected at this level but sentence must make sense to be acceptable.
(BB) – PHRASES: Memorized expressions, phrases (with connection to the verb), or one sentence type.	May make frequent errors, but usually comprehensible to a sympathetic reader/listener. L1 influence may be present.
(BA) – WORDS: A few isolated words, lists of words with no grammatical connection.	Limited language control, inability to create more than individual words. L1 influence may be strong. Errors expected at this level, but must be able to produce at least 2 comprehensible words.
NON-RATABLE: No written or spoken language, non-target language, gibberish, profane/violent language.	NON-RATABLE: No written or spoken language, non-target language, gibberish, profane/violent language.

Table 7

Scores and Proficiency Levels

Score	Level
EB/C	Expanding Mid/High
EA	Expanding Low
TC	Transitioning High
TB	Transitioning Mid
TA	Transitioning Low
BC	Beginning High
BB	Beginning Mid
BA	Beginning Low

References

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

Child, J. R. (1987). Language proficiency levels and the typology of texts. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations, and concepts* (pp. 97–106). Lincolnwood, IL: National Textbook Company.

Child, J. R. (1998). Language skill levels, textual modes, and the rating process. *Foreign Language Annals*, 31(3), 381–391.

Kaftandjieva, F. (2009). Basket procedure: The breadbasket or the basket case of standard setting methods? In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives*. Council of Europe: Cito, Institute for Educational Measurement.

Linacre, J. M. (2008). *Winsteps: A Rasch analysis computer program*. [Version 3.68]. Chicago, IL. (http://www.winsteps.com)

Luecht, R. M. (2003). *Exposure control using adaptive multi-stage item bundles*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, April 22-24, 2003. Chicago, IL.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, 19(3), 189–202.

Rasch, G. (1960/1980). *Probabalistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago.

A First Floor Algorithm

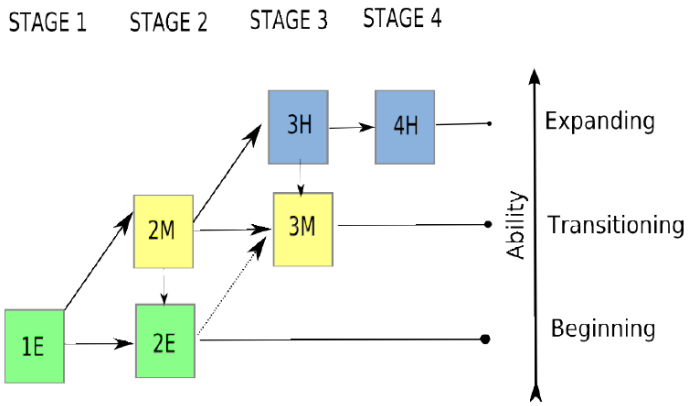

Figure A.1. "Floor first" delivery algorithm used for pilot

B Sample Chinese Benchmark

Literacy Benchmark III
(Based on ACTFL Novice-High)

On all topics from the previous Benchmarks plus:

- Leisure
- Animals
- Health
- Customs/celebrations
- Transportation
- Travel

Students are able to:	With text type:	At performance level:
Communication *Interpretive* • Extract details • Skim for gist	• Weather report • Tickets • Invitations • Notices • Brochures • Instructions • Maps	90% accuracy
Presentational • Express meaning		Comprehensible to a sympathetic reader accustomed to emergent writing

Figure B.1. Chinese Reading Benchmark

C Chinese Reading Crossplot

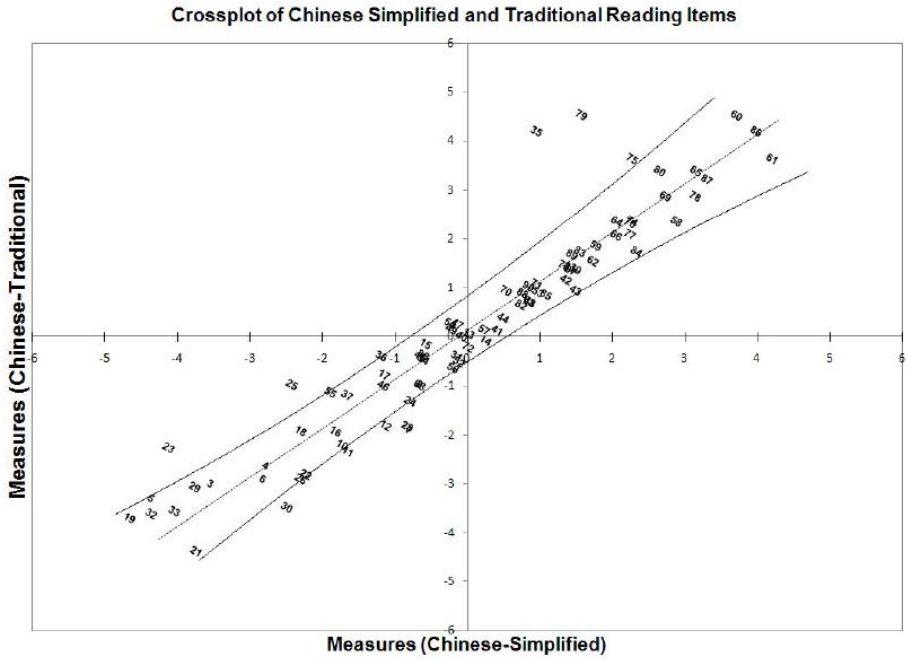

Figure C.1. Crossplot of reading items from traditional and simplified version

D Rasch summary statistics

Table D.1
Chinese Reading Results - Persons

Summary of 731 Measured (Non-Extreme) Persons

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	17.7	26.9	-.58	.53	1.00	.0	.99	.1
S.D.	10.8	12.0	2.08	.16	.21	.9	.56	.8
Max	46.0	47.0	5.45	1.64	1.87	2.5	9.90	6.2
Min	1.0	2.0	-4.91	.35	.33	-2.6	.16	-2.4

Note. Winsteps v3.68 Table 3.1., Real RMSE=.58, Adj.SD=2.00, Separation=3.43, Person Reliability=.92, Model RMSE=.56, Adj.SD=2.01, Separation=3.60, Person Reliability=.93

Table D.2
Chinese Reading Results - Items

Summary of 70 Measured (Non-Extreme) Items

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	188.3	284.6	.00	.17	.98	-.2	.97	-.2
S.D.	148.0	191.4	1.92	.05	.13	1.9	.22	1.6
Max	669.0	742.0	3.66	.35	1.40	6.7	1.60	5.3
Min	30.0	86.0	-4.14	.10	.68	-4.2	.51	-2.7

Note. Winsteps v3.68 Table 3.1., Real RMSE=.18, Adj.SD=1.91, Separation=10.43, Item Reliability=.99, Model RMSE=.18, Adj.SD=1.91, Separation=10.60, Item Reliability=.99

Table D.3
Chinese Listening Results - Persons

Summary of 669 Measured Persons

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	27.8	39.7	.36	.47	1.00	.0	.96	.1
S.D.	13.2	12.4	2.33	.14	.19	.8	.51	.8
Max	54.0	57.0	6.73	1.42	1.78	3.1	4.58	3.2
Min	1.0	2.0	-4.54	.32	.26	-2.5	.03	-2.3

Note. Winsteps v3.68 Table 3.1., Real RMSE=.51, Adj.SD=2.27, Separation=4.42, Person Reliability=.95, Model RMSE=.49, Adj.SD=2.27, Separation=4.62, Person Reliability=.96

Table D.4
Chinese Listening Results - Items

Summary of 85 Measured (Non-Extreme) Items

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	220.1	313.7	.00	.17	.99	-.1	.94	-.3
S.D.	169.7	186.5	2.24	.06	.12	1.8	.27	1.5
Max	607.0	667.0	5.61	.47	1.34	4.5	1.77	3.8
Min	19.0	102.0	-3.68	.12	.76	-3.6	.35	-3.6

Note. Winsteps v3.68 Table 3.1., Real RMSE=.19, Adj.SD=2.24, Separation=12.00, Item Reliability=.99, Model RMSE=.18, Adj.SD=2.24, Separation=12.18, Item Reliability=.99

E Simulation Plot

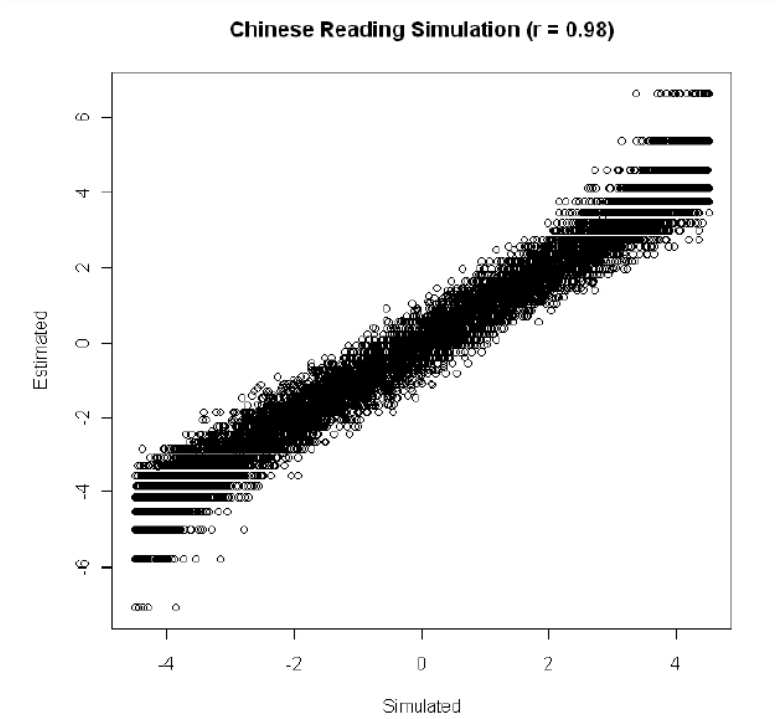

Figure E.1. Simulated ability versus estimated ability correlation

F Standards Setting Agenda

Day 1 (Sunday, August 10) McKenzie 175 ground floor

Agenda Item	Goal	Comments / Technology	Time
Pick up at Secret Garden – walk to Campus	Six rooms (double up)	Kun and Weijun meet guests in lobby	12:15
Kick-off	Introduce participants	Take care of any outstanding paperwork needs Hors d'Oeuvres	12:30
Standard setting intro	Give participants overview of process		1:00
CAP Overview a. Purpose b. Levels c. Benchmarks d. Sample Items e. Algorithm	Give participants an overview of the purpose of CAP, how the test will be used, what the format of items is, and how it will be delivered	Highlight proficiency versus achievement; mid-project changes vis-à-vis STAMP 2.0 project; ACTFL/ILR guidelines intro, vocab at lower levels issue	1:15
Questions / clarifications	Make sure any initial concerns/questions are addressed		1:50
Break		Snacks	2:00
Scale Refresher	Give participants refresher course in scale history, use		2:15
Passage Rating 1	Familiarize participants with passage levels		2:30
Break			3:30
Passage Rating 2	Continue familiarization		3:45
Passage Rating 3	Language Specific	Participants look at Chinese passages and discuss features based on passage rating ideas	4:45
Wrap up	Assign homework		5:15
Finish			5:30
Dinner with Director		Location **McMenamins**	6:30

Figure F.1. Standards setting day one

Day 2 (Monday, August 11) Pacific 113

Breaks & Meals in Pacific 115

Agenda Item	Goal	Comments/ Technology	Time
Breakfast @ 7:45 Pick up at Secret Garden – Walk to campus		Meet in lobby	8:15
Homework Review	Make sure participants have understood passage levels	Set up Turning Point	8:30
Standard setting intro	Process overview, how to log into CAP system		9:00
Round 1 (sample set)	Participants work individually on Round 1 (small set)		8:45
Round 1 discussion	Make sure any initial concerns/questions are addressed	Make sure explicit notes are taken	9:30
Break		Snacks in Pacific 115	10:30
Round 2 rating	Participants work individually on Round 2 items		10:45
Round 2 discussion	Highlight problematic items	Make sure explicit notes are taken	11:45
Lunch		Catering set up in Pacific 115	12:15
Round 3 rating	Participants work individually		1:15
Round 3 discussion	Highlight problematic items	Make sure notes are taken	2:15
Break		Snacks in Pacific 115	2:45
Round 4	Participants work individually		3:00
Round 4 discussion	Highlight problematic items		3:45
Round 5 rating			4:15
Round 5 discussion	Highlight problematic items		5:00
Wrap up		Free Night	5:30

Figure F.2. Standards setting day two

Day 3 (Tuesday, August 12) McKenzie 175 ground floor

Agenda Item	Goal	Comments	Time
Breakfast @ 7:45 Pick up at Secret Garden – Walk to campus		Meet in lobby	8:00
Speaking scale revisited	Give participants explanation of speaking portion of test	Participants can take respond to several prompts	8:10
Speaking samples	Have participants listen to speaking samples		9:00
Speaking prompt review	Have participants rate speaking prompts		10:00
Break		Snacks	10:30
Writing scale	Talk about writing section of test	Participants can respond to several prompts	10:45
Writing samples	Participants see student writing samples and discuss		11:15
Lunch			12:00
Prompt review	Have participants review writing prompts		1:00
Final report	Participants meet to discuss the test and provide list of revisions, changes that they would like to see		2:00
Final report presentation	Give CASLS' staff gist of discussion		3:30
Finish			4:00

Figure F.3. Standards setting day three

G Student survey

G.1 Reading

How would you describe the items on the test?

Generally the right level, but sometimes too easy	45	(5.84%)
Generally the right level, but sometimes too hard	386	(50.06%)
Generally the right level	88	(11.41%)
Too easy for me	23	(2.98%)
Too hard for me	7	(29.70%)
(Blank)	1	(0.13%)

The text for the reading section was clearly legible.

Completely	249	(32.25%)
Mostly	333	(43.13%)
Sometimes	120	(15.54%)
Infrequently	28	(3.63%)
Not at all	42	(5.44%)
(Blank)	0	(0.00%)

The situations used in this test were familiar and easy to understand.

Completely	94	(12.18%)
Mostly	257	(33.29%)
Sometimes	32	(28.32%)
Infrequently	70	(9.07%)
Not at all	68	(8.81%)
(Blank)	0	(0.00%)

How would you describe the length of time it took to complete this test?

Just right	383	(49.61%)
Too long	365	(47.28%)
Too short	24	(3.11%)
(Blank)	0	(0.00%)

Overall, how appropriate do you think a test like this is for measuring your ability to read Chinese?

Very appropriate	35	(30.97%)
Somewhat appropriate	442	(57.0%)
Not very appropriate	85	(11.01%)
Not at all appropriate	56	(7.25%)
(Blank)	2	(1.77%)

G.2 Listening

How would you describe the items on the test?

Generally the right level, but sometimes too easy	45	(6.82%)
Generally the right level, but sometimes too hard	337	(51.06%)
Generally the right level	112	(16.97%)
Too easy for me	16	(2.42%)
Too hard for me	150	(22.73%)
(Blank)	3	(3.53%)

How would you describe the audio on this test?

All very clear and audible	138	(20.88%)
Mostly clear and audible	291	(44.02%)
Some were unclear or volume inadequate	173	(26.17%)
Many were unclear or had poor volume	59	(8.93%)

The situations used in this test were familiar and easy to understand.

Completely	65	(9.83%)
Mostly	301	(45.54%)
Sometimes	198	(29.95%)
Infrequently	51	(7.72%)
Not at all	46	(6.96%)

How would you describe the length of time it took to complete this test?

Just right	346	(52.45%)
Too long	293	(44.39%)
Too short	21	(3.18%)

Overall, how appropriate do you think a test like this is for measuring your ability to understand spoken Chinese?

Very appropriate	223	(33.89%)
Somewhat appropriate	349	(53.04%)
Not very appropriate	48	(7.29%)
Not at all appropriate	38	(5.78%)