STAMP™ 4Se (STAndards-based Measurement of Proficiency – 4 Skills Elementary)

Technical Report

By Martyn Clark and Linda Forrest

Center for Applied Second Language Studies (CASLS)

Updated by Dr. Jim Snyder, Director of Market Research

Avant Assessment LLC

10/4/2012

This project was supported by grants from the U.S. Department of Education and the Oregon Chinese Flagship program at the University of Oregon.

This report is an updated version of a previous report titled *National Online Early Language Learning Assessment (NOELLA)* published by CASLS (June 2009 http://casls.uoregon.edu/papers.php). The writing and listening scoring section was updated to reflect Avant Assessment's process and the previous report included some additional test functionality and languages that were not included in the current STAMP 4Se delivered by Avant Assessment.

Introduction

CASLS is a Title VI K-16 National Foreign Language Resource Center located at the University of Oregon. CASLS' core mission is promoting international literacy by supporting communities of educators and by partnering with those communities to develop a comprehensive system of proficiency-based tools for lifelong language learning and teaching.

CASLS is supported almost exclusively by 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:

- Teachers are the solution, not the problem. Support them, don't preach to them.
- All children have the ability to learn a second language and should be provided with that opportunity.
- The purpose of language learning is meaningful communication.
- Meeting the needs of teachers and students is our top priority.

The STAMP 4Se is a web-based test of general proficiency for children learning foreign languages in Grade 3 through Grade 6. It is based, in part, on the Standards-based Measurement of Proficiency (STAMP), which was created by CASLS to assess the language proficiency of students age 13 and above. CASLS has a technology transfer agreement with Avant Assessment to license CASLS assessments, and STAMP is currently supported and delivered by Avant.

STAMP 4Se has been developed in seven languages: Spanish, French, Japanese, and Chinese, Hebrew, Russian, and Korean. Development of the assessment for these languages has been funded from a variety of sources and in collaboration with a number of partners. The STAMP 4Se project was initially funded by a Foreign Language Assistance Program (FLAP) grant to the state of Wyoming. Wyoming, acting as part of a consortium of six states (including South Carolina, New Jersey, Georgia, Kentucky, and Virginia), sponsored development of Spanish, Japanese, and French assessments for Benchmark levels 1 through 4. CASLS supported the development of items at Benchmarks 5 and 6 using National Foreign Language Resource Center funding. Concurrently, the University of Oregon Chinese Flagship sponsored the development of a Chinese version of STAMP 4Se. Additional funding was provided by a FLAP grant to the state of Georgia to develop teacher reporting pages for all languages.

Content for STAMP 4Se was developed by CASLS working with the Wyoming Department of Education in collaboration with the twenty-six (26) elementary schools in the state's K-6 language programs and elementary schools in the cooperating states. The Center for Applied Linguistics (CAL) in Washington, D.C., worked with these partners to develop Spanish STAMP 4Se. The Oregon Chinese Flagship Program provided resources and personnel to develop Chinese STAMP 4Se.

Description of the assessment

STAMP 4Se is a web-based test of proficiency, with versions available for Spanish, French, Japanese, and Chinese. It is appropriate for upper elementary learners studying those languages as a second or foreign language (Grade 3 through Grade 6.) The test is based on Benchmark Specifications developed jointly by CASLS, CAL, and language-specific teams of elementary school immersion and FLES teachers. These Benchmarks correlate closely with the ACTFL K-12 Performance Guidelines in the Novice and Intermediate range. STAMP 4Se is designed to evaluate general language proficiency. As such, it is not based on any specific syllabus or teaching program. Test scores are reported on the CASLS' Benchmark scale and the closest ACTFL scale correlate.

Content and structure of STAMP 4Se

Test items are situated within the context of daily school life. The students mentioned in items are attending an elementary school in the U.S. (Parkhurst Elementary). In this school, teachers and students often speak in the target language. Some are native speakers of the language, while others are native English speakers. Two children who are native speakers of the target language, one boy and one girl, are introduced at the beginning of the test. Test takers learn that they have recently come to Parkhurst from another country.

STAMP 4Se consists of four sections:

1. Interpretive Listening
2. Interpretive Reading
3. Presentational Writing
4. Presentational Speaking

Each of these sections is described below.

Interpretive Listening

The interpretive listening section consists of a series of dialogues and monologues in the target language. Each dialogue or monologue is followed by a question in the target language. The passage and question are heard twice. Students indicate the correct answer either by clicking on the correct picture in a set of four pictures (picture selection) or by clicking on the relevant area in a single picture (picture click). The questions assess the test-taker's ability to understand the gist of the passage as well as to extract detailed information. The dialogues and monologues are all performed by fluent speakers of the target language and are delivered at an age-appropriate speed. The listening section is presented adaptively. After each group of ten items, the computer chooses the next group to be administered.

Reading Comprehension

The reading comprehension section is a multiple-choice test designed to evaluate the reader's ability to scan written passages for gist and to extract detailed information. All of the passages are designed to mimic authentic reading tasks, such as reading signs, journal entries, or classroom materials. The reading passages are of a general nature and do not assume specialized knowledge of culture or customs. Students indicate the correct answer either by clicking on the correct picture in a set of four pictures (picture selection) or by clicking on the relevant area in a single picture (picture click). In a third type of item, students view a picture and read the question and answer choices in the target language. The reading section is also presented adaptively.

Presentational Writing

The presentational writing section tests the test-takers' ability to express themselves in the target language through two short writing tasks. The writing tasks are presented aurally in English. Following the task description, test takers are reminded aurally in the target language to 'remember to write in <*language name*>'. Test takers respond to the tasks in the target language by typing their answers directly into the computer. Though the writing section is computer delivered, it is not adaptive. The written responses are graded by teachers according to a simple rubric. Note that this section assumes that test-takers have familiarity with keyboarding in the target language.

Presentational Speaking

The presentational speaking section tests the examinee's ability to express themselves in the spoken language through two short tasks. The speaking tasks are non-interactive (i.e., not an interview or conversation). The speaking tasks are presented aurally in English. Following the task description, test takers are reminded aurally in the target language to 'remember to speak in <*language name*>'.Test takers record their responses directly into the computer using a microphone. The responses are graded by teachers according to a simple rubric.

Description of the test taker

The target audience for this test is students in Grade 3 through Grade 6 studying foreign languages. The test takers will most likely be students in FLES or immersion programs. STAMP 4Se items are designed to assessment students whose proficiency levels fall with CASLS' Benchmark levels 1 through 6, which correlate to ACTFL levels Novice-Low through Intermediate-High. Consequently, the test may not accurately measure the language proficiency of some heritage or immersion program students.

Literacy in English is not assumed in STAMP 4Se, except for familiarity with numerical digits. Any important written material occurs only in the target language and then only in the reading assessment; any other writing that occurs within the images is considered decorative in nature and not necessary for correctly responding to the item (e.g. a sign on a building in a picture). Instructions are always provided aurally. All critical instructions are given in English. Other instructions, such as cues to choose the correct answer or to speak or write in the target language, are given in the target language.

Description of the test score user

Students, language instructors, parents, and program administrators are the intended score users. Scores are reported by class to the classroom teacher, and it is assumed that other potential test score users will receive the score from the teacher. Students will use the test score to evaluate their progress towards their language learning goals. Language instructors will use the scores to help inform (in conjunction with multiple other sources of information) summative evaluations of the students and class progress. At the class level, aggregate information can help inform curricular decisions for educators and program administrators.

Intended consequences of test score use

STAMP 4Se is intended to improve language teaching and learning by providing information on student proficiency. The goal of providing this information is to create positive washback between the test and the language program. STAMP 4Se scores should not be used for punitive purposes.

Limitations on score interpretation

As with any test, STAMP 4Se scores should be considered one piece of evidence for a child's proficiency. Students, especially young students, can perform differently on different days due to a variety of factors. STAMP 4Se is designed to give a general snapshot of proficiency with a fairly limited number of items. STAMP 4Se scores should not be used for high-stakes decisions, such as final grades or exit exams.

Construct for STAMP 4Se

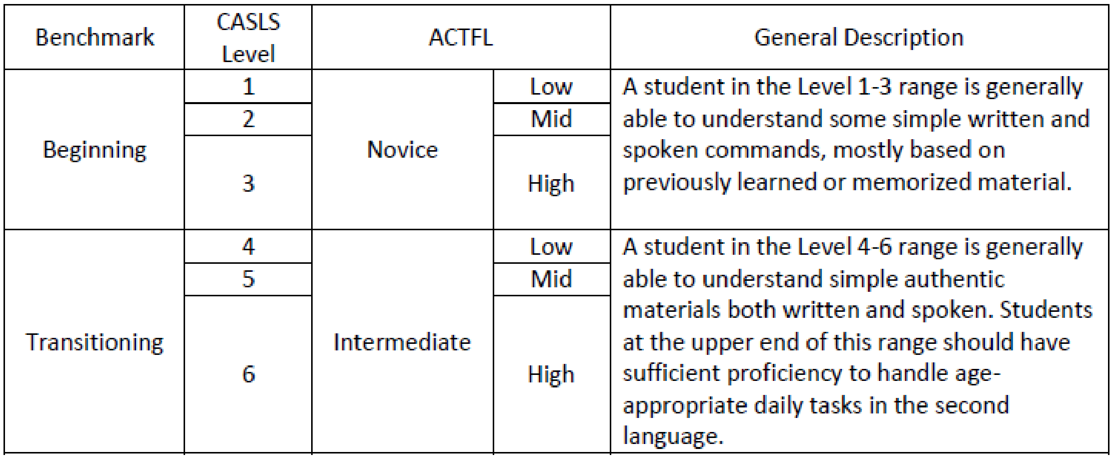
STAMP 4Se can be considered a "proficiency-oriented" test. Language proficiency is a measure of a person's ability to use a given language to convey and comprehend meaningful content in realistic situations. STAMP 4Se is intended to gauge a student's linguistic capacity for successfully performing language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

STAMP 4Se results are based on Benchmarks that describe the language ability expected at different proficiency levels. As such, the Benchmarks can be considered a general hypothesis about proficiency at these levels. During the test development phase (described later), empirical evidence for these hypotheses was collected and incorporated into the final test.

Test level

STAMP 4Se is designed to assess students with proficiency levels in the range of 1 through 6 on the CASLS Benchmark Scale. Several points along this scale have been designated as Benchmark levels. These Benchmark Levels include verbal descriptions of the proficiency profile of the typical student at that point in the scale.

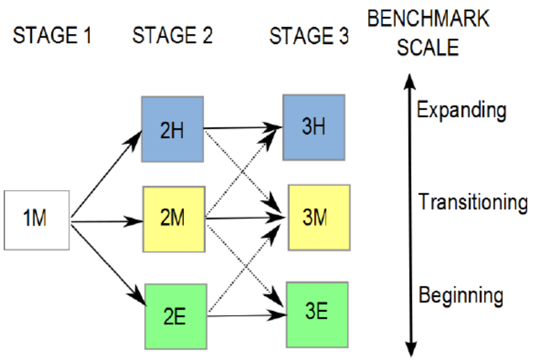
The Benchmark Level descriptions are intended to be comparable to well-known proficiency scales, notably the ACTFL Proficiency Guidelines. The relationship between the scales is shown in the table below.



Benchmark	CASLS Level	ACTFL		General Description
Beginning	1	Novice	Low	A student in the Level 1-3 range is generally able to understand some simple written and spoken commands, mostly based on previously learned or memorized material.
	2		Mid	
	3		High	
Transitioning	4	Intermediate	Low	A student in the Level 4-6 range is generally able to understand simple authentic materials both written and spoken. Students at the upper end of this range should have sufficient proficiency to handle age-appropriate daily tasks in the second language.
	5		Mid	
	6		High	

Test Delivery

STAMP 4Se is delivered over the internet using a standard web browser. Logins for the test are created at the class, not individual, level. It is expected that the test will be delivered in a proctored environment, such as a school's computer lab. The reading and listening sections of STAMP 4Se were designed to be delivered using a multistage adaptive algorithm. Items in the test are arranged into multi-item testlets or bins of different difficulty. 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 or easier. This algorithm is illustrated graphically below.



Multistage adaptive algorithm

Test Development

Benchmark Development

Benchmarks for each of the languages were developed by committees of foreign language educators in a series of workshops. Two separate workshops were held: one to develop French and Spanish benchmarks and one to develop Japanese and Chinese benchmarks. Workshop attendees were educators nominated by the cooperating states or involved in elementary foreign language programs in other areas of the U.S., along with CASLS staff, grant PIs, and representatives from the partnering organizations.

Each benchmark group was given an overview of the project and sample benchmarks from the STAMP, CASLS' proficiency test for students age 13 and over. The committee was instructed to create benchmarks for all four skills that would be appropriate for elementary school children while being consistent with the ACTFL K-12 Proficiency Guidelines and the National Standards. The Benchmark for each level would contain detailed age-appropriate specifications in relation to the topics and functions expected of language learners at each proficiency level. Sample Benchmarks and a complete list of workshop dates and participants can be found in the Appendix.

Item Development

As with Benchmark development, items were also initially developed by groups of foreign language educators in a series of workshops. For each workshop, participants were given an overview of the test and the Benchmarks. Next, basic item writing guidelines were presented. Finally, the participants were divided into language-specific groups for the actual item writing. CASLS staff members were present in each of the item writing groups. After the item writing sessions, CASLS staff reviewed items and chose the most promising for further development. Appropriate graphics and audio files were created, reviewed, and modified if necessary, over a period of several months. Once completed, these were uploaded into the delivery system and reviewed again. Concurrent with item development, the technical infrastructure of the CASLS test delivery system was updated to include the new item types and delivery engine that STAMP 4Se required. The programming and testing of these features continued throughout the project.

Empirical Validation

Pre-pilot Testing

Concerns had been raised during the Benchmark development phase about the use of English versus the target language on the test. Some immersion instructors felt that the entire test should be in the target language while others feared that students would not understand instructions not given in English. To investigate this issue as well as try out some of the new technical features of the delivery engine, a pre-pilot was conducted using some completed Spanish items. Two versions of each item in the pilot were created, one with instructions in English and one with instructions in Spanish. The results indicated that there was no detrimental effect for presenting the instructions in Spanish, and it was decided to present the instructions in the target language for all STAMP 4Se versions.[1]

Pilot Testing

Items created for STAMP 4Se were piloted as "fixed-form" tests to collect empirical data on the functioning of the items. This piloting was done in two stages, with one pilot starting in fall 2006 and the second pilot in spring 2007. Five test forms consisting of items from adjacent proficiency levels were created. These were piloted in participating schools nationwide. CASLS staff also visited several local schools to observe students taking the test.

After each pilot, a Rasch analysis was conducted on the reading and listening data. Items that were not behaving as expected were revised or discarded. Speaking and writing samples were collected during piloting for later use as training samples. Over 7,000 tests were delivered during the pilot phase.

Field Testing

Items successfully passing the first two rounds of pilot testing were chosen for field testing. Two field tests were conducted, the first in fall 2007 and the second in spring 2008. The field tests were delivered adaptively using the finalized multistage delivery engine. These field tests were intended to ensure that the delivery algorithm was working properly and that the test would be ready for operation. In addition, score reporting by class was implemented for the field tests. Results from the field test indicated that the multistage algorithm was working appropriately. Slight changes were made between the first and second field test to finalize the bin sizes for the delivery algorithm. Over 13,000 tests were delivered during the field testing phase.

Scaling

The field test data from the finalized versions of the items was used to scale the test for scoring purposes. Items were scaled using Rasch analysis, and cut points were set on the Rasch scale. Cut points were set for the ability level at which a test-taker has an 80% chance of being correct on an item of median difficulty for the level in question. The proficiency rating that the student receives at the end of the test is taken from a scoring table that considers their test path (i.e., the particular items that they took on the test)

[1] The instructions here refer to the specific instructions for each item, not the general instructions at the beginning of the test. Those initial instructions are presented in English.

and their total score. Thus, students with the same total score may get different proficiency ratings if one of the students took a test of more difficult items. Simulation studies with the finalized items and score tables indicate that the students are classified within ± 1 level of their “true” ability level approximately 98% of the time.

Current Status

As of this updated report (October 2012), Spanish, French, Chinese and Japanese STAMP 4Se have been finalized and are currently available through Avant Assessment’s website.

Appendix 1 - Sample Benchmarks (French)

Reading Benchmark I – grades 3-6

(based on ACTFL Novice Low)

Function	**Context/Text Type**	**Performance Level**
students should be able to …	in …	by . . .
• identify cognates • identify common words in context	• signs (traffic, commercial) • lists of words • high frequency phrases	

Reading Benchmark II – grades 3-6

(based on ACTFL Novice Mid)

Function	**Context/Text Type**	**Performance Level**
students should be able to …	in …	by . . .
• identify information • derive meaning	• advertisements • labels • titles (in context – books, poems, songs)	

Reading Benchmark III – grades 3-6

(based on ACTFL Novice High)

Function	**Context/Text Type**	**Performance Level**
students should be able to …	in …	by . . .
• identify information • derive meaning • compare and contrast	• maps • instructions/directions • surveys • charts and graphs	

Reading Benchmark IV – grades 3-6

(based on ACTFL Intermediate Low)

Function	Context/Text Type	Performance Level
students should be able to …	in …	by . . .
• show emerging use of linguistic context to identify the meaning of unfamiliar language • skim for gist • identify the main idea	• simple narratives (stories) • invitations (birthdays, holiday celebrations) • simple descriptions • simple poems and rhymes	

Reading Benchmark V – grades 3-6

(based on ACTFL Intermediate Mid)

Function	Context/Text Type	Performance Level
students should be able to …	in …	by . . .
	• short children's literature below L1 reading level • simple non-fiction texts on familiar subjects (textbooks, children's magazines) • games and puzzles	

Reading Benchmark VI – grades 3-6

(based on ACTFL Intermediate High)

Function	Context/Text Type	Performance Level
students should be able to …	in …	by . . .
• infer meaning based on overall comprehension of a reading	• multi-paragraph fiction and non-fiction texts	

Appendix 2 – Benchmark Workshops

Workshop 1 - French, Spanish

Location : Richmond, VA
Dates : October 16 – 17, 2005
Participants: Alison Moran, Alicia Vinson, Elsa Batista, Dawn Samples, Kathy Duran, Cassandra Celaya

Workshop 2 – Chinese, Japanese

Location : Portland, OR
Dates: May 12, 2006
Participants : Shuhan Wang, Yu-Lan Lin, Jessica Bucknam, Atsuko Ando, Hiroko Darnell, Lynn Sessler, Jennifer Pedersen

Appendix 3 – Item Writing Workshops

Workshop 1 – Spanish

Location : Washington, DC
Dates : January 16 – 18, 2006
Participants : Alicia Vinson, Marci Bland, Elsa Batista, Stephanie Cano, Mark Eastburn, Dawn Samples, Lynn Fulton-Archer, Gloria Quave, Mary Eileen Yaeger, Kathy Duran, Angelica Echevarria, Luisa Sanchez

Workshop 2 – Chinese, French, Japanese

Location : Portland, OR
Dates : June 21 – 23, 2006
Participants : Hiroko Darnell, Jennifer Pedersen, Kayo Imamura, Kayoko Kasai, Lili Kennington, Masakazu Yamakawa, Matt Bacon-Brenes, Michiko Parshalle, Mieko Imanishi, Miho Nakagawa, Naomi Hashimoto Kraft, Yoshiko Kamata, Adrianne Bee, Beimei Long, Catherine Huang, Chusheng Tang Liao, Cindy Lin, Jessica Bucknam, Jiun Chou Young, Kit Nadeau, Liduan Hugel, Linda Tong, Mary Jew, Shen Ying, Xiaoping Xie, Alison Moran, Annie Dwyer, Dawn Samples, Evangeline Reddick, Jean Amick, Jennifer Bernhard, Joelle Chivers, Julie Riggs, Leslie Vandeventer, Paola Durant, Stephanie Appel

Appendix 4 –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 STAMP4S rubric is as follows:

Table 1 STAMP 4Se 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 2

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