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

Technical Report

Updated by Victor D.O. Santos, Ph.D.

Avant Assessment

12/01/2023

Introduction

The STAMP 4Se is a multistage, computer-adaptive test of real-world, general language proficiency for children learning foreign languages in Grade 3 through Grade 6, initially developed by CASLS (Center for Advanced Second Language Studies) at the University of Oregon, with support from the U.S. Department of Education. It is based, in part, on the Standards-based Measurement of Proficiency (STAMP 4S), which was created by CASLS to assess the language proficiency of students aged 13 and above. Both STAMP 4S and STAMP 4Se are currently fully developed, supported, and delivered by Avant Assessment.

STAMP 4Se has been developed in fifteen languages: Arabic, Cantonese, English, French, German, Hawaiian (ʻŌlelo Hawaiʻi), Hebrew, Japanese, Korean, Mandarin Simplified, Mandarin Traditional, Portuguese (Brazilian), Russian, Spanish, and Yupʼik. Development of the assessment for the initial languages – Chinese (Simplified and Traditional), French, Hebrew, Japanese, Korean, Russian, and Spanish – was funded from a variety of sources and in collaboration with several 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 levels 1 (Novice Low) through 4 (Intermediate Low). CASLS supported the development of items at levels 5 (Intermediate Mid) and 6 (Intermediate High) using National Foreign Language Resource Center funding. 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. For the remaining languages – Arabic, Cantonese, English, German, Hawaiian (ʻŌlelo Hawaiʻi), Portuguese (Brazilian), and Yupʼik – all development has been done internally by Avant.

Content for STAMP 4Se was initially developed by CASLS working with the Wyoming Department of Education in collaboration with the 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. For all other languages, development has been done by a group of language-assessment experts at Avant Assessment, which includes target-language-experts in each language.

Description of the Assessment

STAMP 4Se is a multistage computer-adaptive test of general language proficiency aligned to the ACTFL proficiency scale and currently available in 15 languages. It is appropriate for upper elementary learners studying those languages as a second or foreign language (Grade 3 through Grade 6) and assesses general, real-world Reading, Writing, Listening, and Speaking proficiency. As such, it is not based on any specific syllabus or teaching program. The test is based on benchmark specifications developed jointly by CASLS, CAL, and language-specific teams of elementary school immersion and FLES teachers. Test scores are reported on the CASLS/STAMP scale, with scores ranging from 1 (Novice Low) to STAMP 6[1] (Intermediate High) for Reading and Listening and from a STAMP 1 (Novice Low) to STAMP 8 (Advanced Mid) for the Writing and Speaking sections.

The Reading and Listening sections of STAMP 4Se are automatically scored and are computer-adaptive, in which the difficulty of the items adapts to the estimated proficiency of the examinee at specific points in the test, as shown in Figure 1:

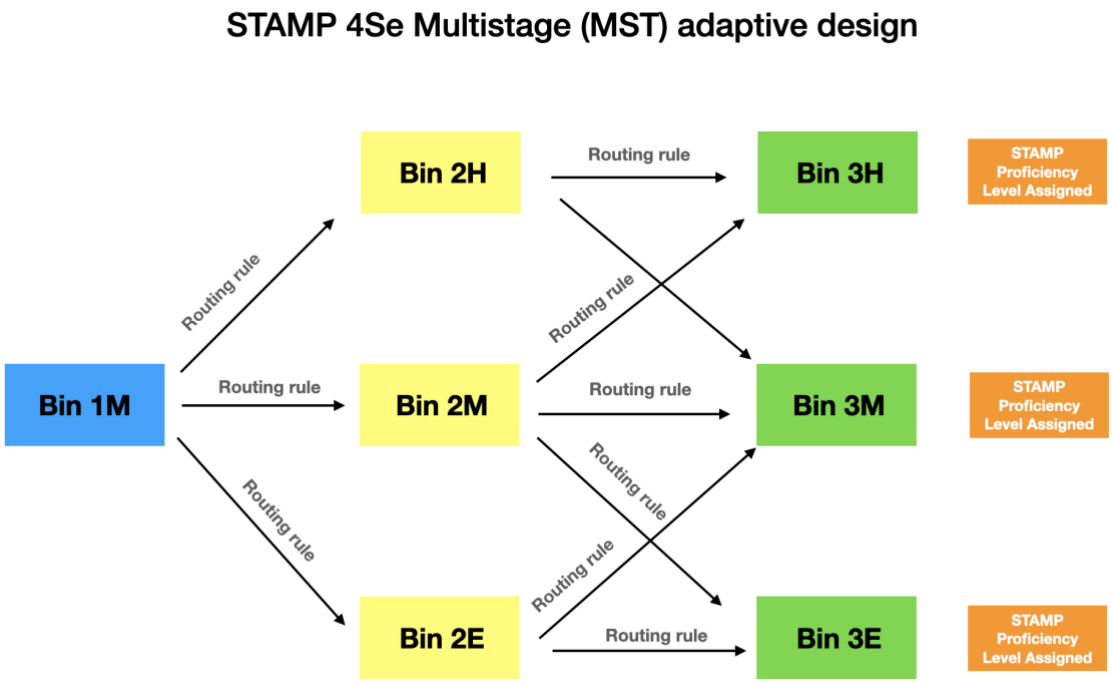

Figure 1. *Multistage adaptive design of the Reading and Listening sections of STAMP 4Se*.

The Writing and Speaking sections are scored by human raters trained on the ACTFL proficiency scale and who must pass a stringent certification process to be allowed to rate live test responses. Each rater is continuously monitored to ensure that their ratings are accurate.

[1] A STAMP level of 6+ is awarded to those examinees who get all items in the test correct, indicating their Reading/Listening proficiency is above Intermediate High.

Content and Structure of STAMP 4Se

Test items are situated within the context of daily school life as well as social and home contexts relevant to students in grades 3-6.

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 by either 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 10 items, the computer chooses the next group of items to be administered, depending on the examinee's demonstrated proficiency up to that specific point in the test.

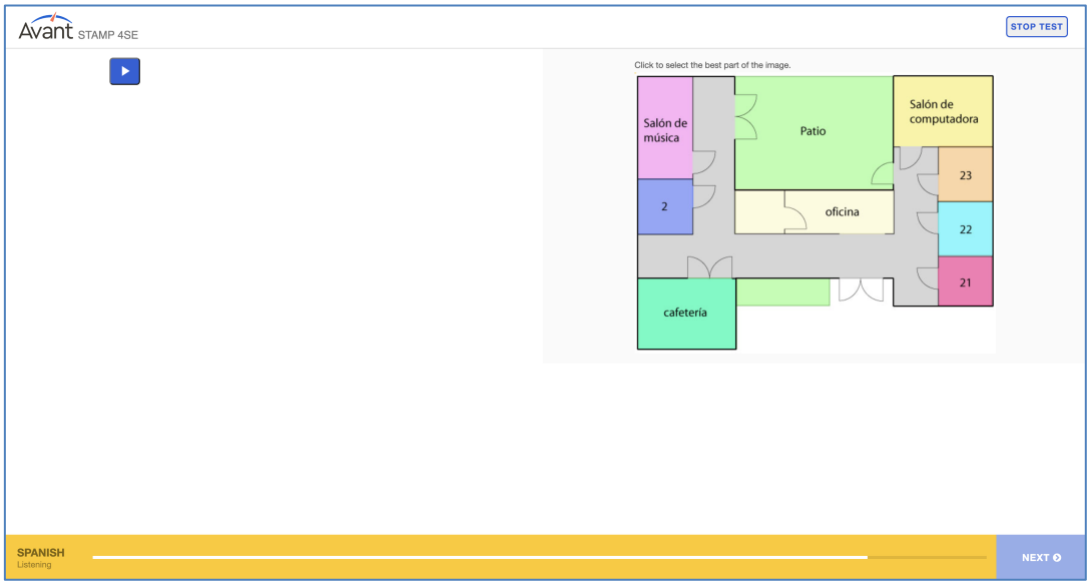

Figure 2. *Sample picture-click item from the Interpretive Listening section of a STAMP 4Se Spanish test.*

Interpretive Reading

The Interpretive Reading section evaluates the examinee's ability to scan written passages for gist and to extract detailed information. All 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 item type (multiple-choice question) students view a picture and read the answer choices and question in the target language. As with the Listening section, the Reading section is also presented adaptively.

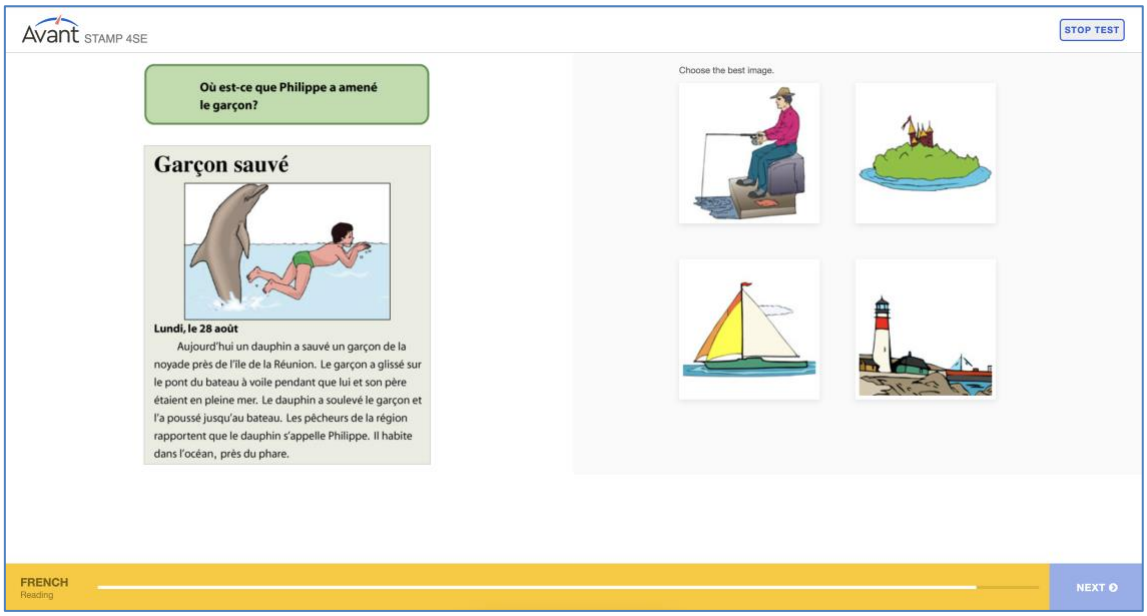

Figure 3. *Picture-selection item from the Interpretive Reading section of a STAMP 4Se French test.*

Presentational Writing

The Presentational Writing section assesses examinees' ability to express themselves in the target language through three writing tasks. The writing tasks are presented textually in English and also aurally in the target language. Following the task description, examinees are reminded to write in the target language, use complete sentences, and write as much as they can, up to a maximum of 2,500 characters allowed by the system. Examinees respond to the tasks in the target language by typing their answers directly into the computer. Although the Presentational Writing section is computer-delivered, it is not adaptive. The written responses are graded by trained human raters who make use of a rating rubric (see Appendix for an overview of the rubric). Note that this section assumes that examinees have familiarity with keyboarding in the target language. Avant also has developed the functionality within our system to have examinees complete the Writing section via a handwritten response as well. This is something that each testing location can request if it is needed for a group of examinees.

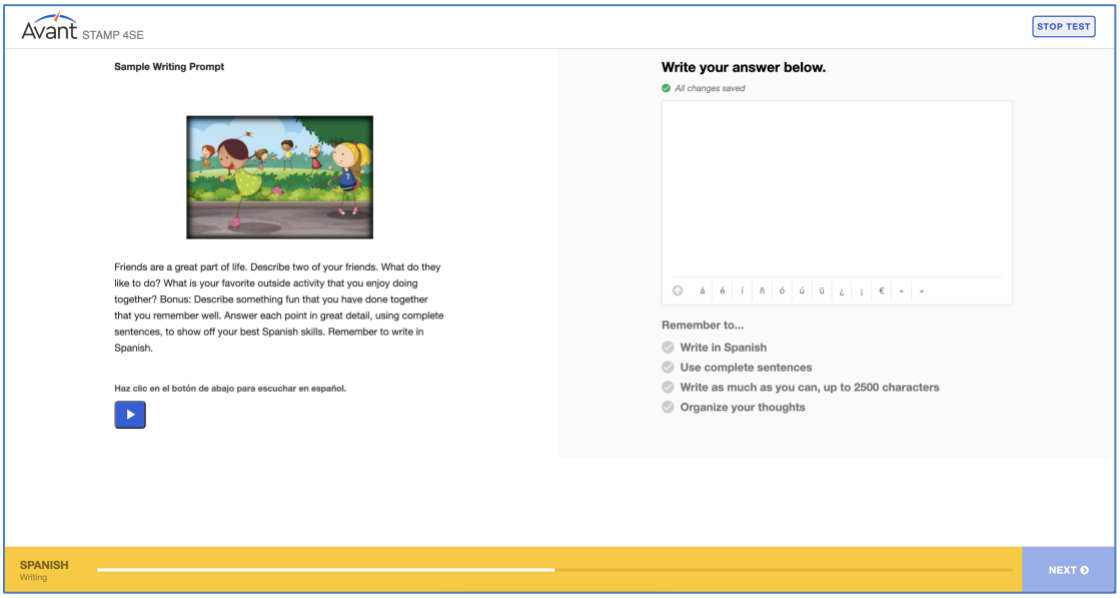

Figure 4. *Sample test in the Presentational Writing section of STAMP 4Se Spanish.*

Presentational Speaking

The Presentational Speaking section assesses examinees' ability to express themselves in the spoken language through three speaking tasks. The speaking tasks are non-interactive (*i.e.*, not an interview or conversation). The tasks are presented textually in English and aurally in the target language. Following the task description, examinees are reminded to speak in the target language, use complete sentences, and speak as much as possible, with a maximum of three minutes allowed by the system. Examinees record their responses directly into the computer using a microphone. Although the Presentational Speaking section is computer-delivered, it is not adaptive. The spoken responses are graded by trained human raters who make use of a rating rubric (see Appendix for rubric).

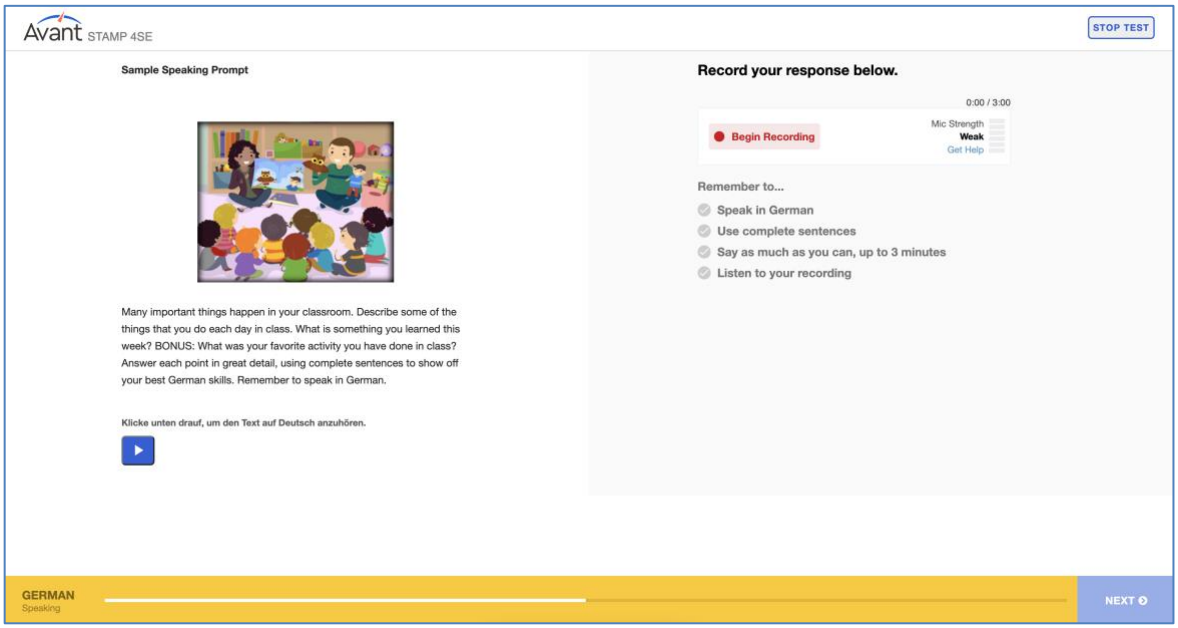

Figure 5. *Sample task in the Presentational Speaking section of STAMP 4Se German.*

Description of the Examinee

The target audience for this test is students in Grade 3 through Grade 6 studying foreign languages. Examinees will most likely be students in FLES or immersion programs. STAMP 4Se items are designed to assess students whose proficiency levels fall within ACTFL proficiency levels Novice Low (CASLS/STAMP level 1) through Intermediate High (CASLS/STAMP level 6). Consequently, the test may not accurately measure the language proficiency of some heritage or immersion program students whose ability may be above Intermediate High.

Literacy or fluency in English is not assumed or required in STAMP 4Se. All Reading and Listening passages, as well as most answer options in the case of multiple-choice questions in these two receptive sections, are provided solely in the target language. For the Writing and Speaking sections, all prompts are provided in written form in English and audio form in the target language. At the beginning of each section, all instructions are provided in written form in both English and 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 or administrator. Students will use the test score to evaluate their progress towards their language learning goals and to identify personal language proficiency strengths and weaknesses. 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 as well as program structure for program administrators. In addition to these test scores, which are available to all clients, more in-depth, customized analyses of the data employing visualization software is also available and may be purchased separately from Avant.

Intended Consequences of Test Score Use and Interpretation

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 a positive washback between the test and the language program. STAMP 4Se scores should not be used for punitive purposes.

As with any test, STAMP 4Se scores should be considered one piece of evidence for a student'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 is 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 scores are based on the language ability expected at different proficiency levels aligned to the ACTFL proficiency scale.

Test Levels

Interpretive Reading and Interpretive Listening

For Interpretive Reading and Interpretive Listening, STAMP 4Se is designed to assess students with proficiency levels in the range of 1 through 6 on the CASLS/STAMP scale. The relationship between the CASLS/STAMP scale and the ACTFL proficiency scale can be seen in Table 1:

CASLS/ STAMP Level	**ACTFL**		**Function** ***(students should be able to)***	**Context /Text Type**
1	Novice	Low	- identify cognates - identify common words in context	- signs (traffic, commercial) - lists of words - high frequency phrases
2		Mid	- identify information - derive meaning	- advertisements - labels - titles (in context – books, poems, songs)
3		High	- identify information - derive meaning - compare and contrast	- maps - instructions/directions - surveys - charts and graphs
4	Intermediate	Low	- show emerging use of linguistic context to identify meaning of unfamiliar language - skim for gist - identify the main idea	- simple narratives (stories) - invitations (birthdays, holiday celebrations, etc.)
5		Mid	- understand the main idea and key information in short, non-complex passages that deal with basic and familiar personal and social topics	- short children's literature below L1 reading level - simple non-fiction texts on familiar subjects (textbooks, children's magazines, etc.)
6		High	- infer meaning based on overall comprehension of a passage and contextual clues - understand some connected passages that feature description and narration across various time frames	- multi-paragraph fiction and non-fiction texts

Table 1. Overview of the CASLS/STAMP scale and its relationship to the ACTFL Proficiency scale

Presentational Writing and Presentational Speaking

For Presentational Writing and Presentational Speaking, STAMP 4Se is designed to assess students with proficiency levels in the range of levels 1 (Novice Low) through 8 (Advanced Mid) on the CASLS/STAMP scale. The relationship between the CASLS/STAMP scale and the ACTFL proficiency scale for these two skills can be seen in Appendix 3.

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 (Figure 1 above). 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 bins. 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.

Initial Test Development and Validation by CASLS

Defining Each Level and its Descriptors

Levels and their descriptors 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 Level descriptions and one to develop Japanese and Chinese Level descriptions. 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 group was given an overview of the project and sample level descriptions from the STAMP 4S, for students aged 13 and over. The committee was instructed to create levels and descriptors 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 levels and their descriptors would contain detailed age-appropriate specifications in relation to the topics and functions expected of language learners at each proficiency level. A complete list of workshop dates and participants can be found in Appendix 1.

Item Development

As with level definition and descriptor development, items were also initially developed by groups of foreign language educators in a series of workshops (Appendix 2). For each workshop, participants were given an overview of the test levels and the level descriptors. 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.

Pre-pilot Testing

Concerns had been raised during the level benchmarking and descriptor 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.

Cut Score Setting

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 specific 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.

Post-CASLS Test Development and Validation

Since the initial test development and validation by CASLS described above, Avant Assessment has internally developed the STAMP 4Se test in several additional languages, while continuing to ensure the high quality of the test and maintain and update the languages initially developed by CASLS. Avant is now able to field-test new items in a much more efficient way, namely by introducing new items as neutral (non-scored) items on live STAMP 4Se tests. After each item has been taken by enough examinees, the item is pulled from the test, analyzed by means of Rasch-measurement and Classical Test Theory methods, and if deemed to meet the STAMP 4Se standards, subsequently introduced into the test as a scored item.

Difficulty Hierarchy of Reading and Listening Items at STAMP levels 1 (Novice Low) through 6 (Intermediate High)

Given that the setting of cut-scores is dependent on a hierarchy of level difficulty and given that the STAMP level awarded to an examinee at the end of their STAMP 4Se administration is dependent on how many items they got correct on the path they followed in their adaptive test as well as on the ACTFL level of the items they encountered on that path, it is crucial to show that the expected difficulty hierarchy of items at the various STAMP/ACTFL levels is observed across the STAMP 4Se in different languages. Below, in Figures 6 to 11, we can see the average Rasch difficulty of items at levels 1 (Novice Low - NL) to 6 (Intermediate High - IH) for the Reading and Listening sections of three STAMP 4Se languages: Spanish, Chinese, and French. The figures show that the expected level hierarchy is preserved for all three sample languages, which provides strong evidence for the quality of the items on STAMP 4Se.

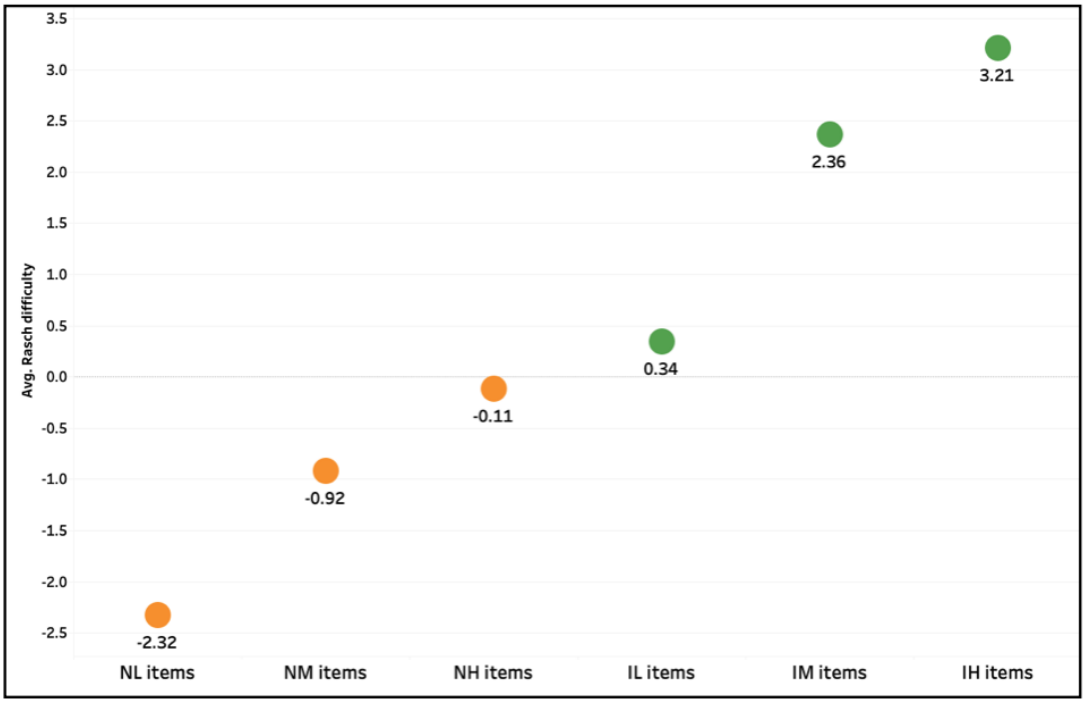

Figure 6. *Average Rasch difficulty of STAMP 4Se Spanish Reading items at levels 1 (Novice Low – NL) through 6 (Intermediate High – IH)*

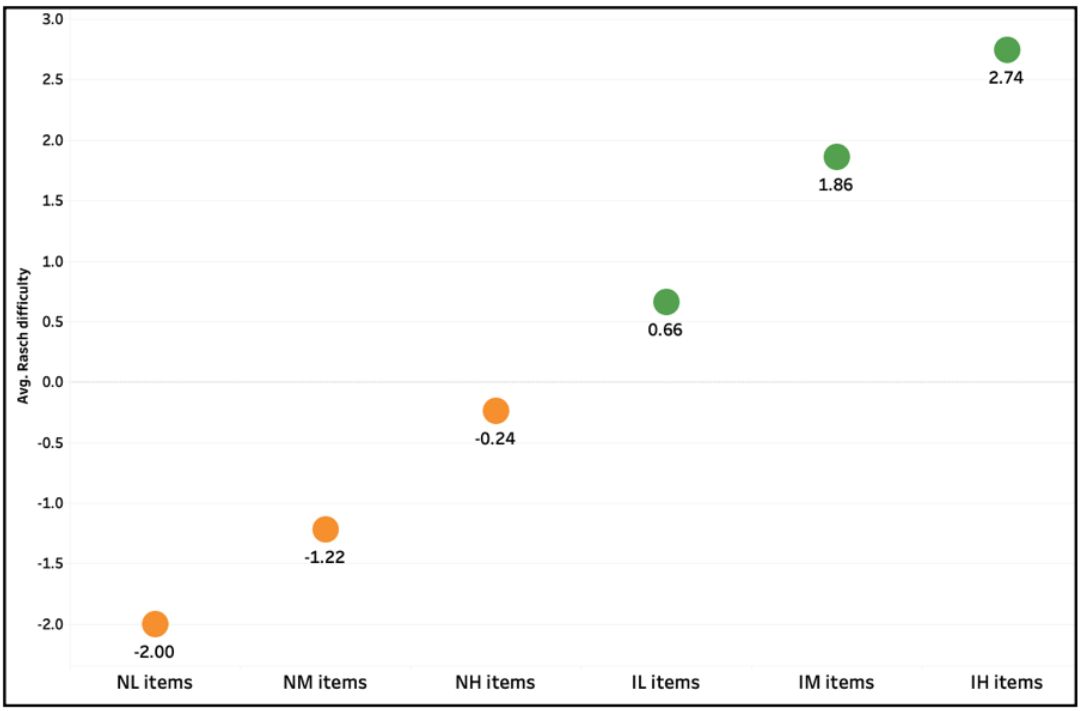

Figure 7. *Average Rasch difficulty of STAMP 4Se Spanish Listening items at levels 1 (Novice Low – NL) through 6 (Intermediate High – IH)*

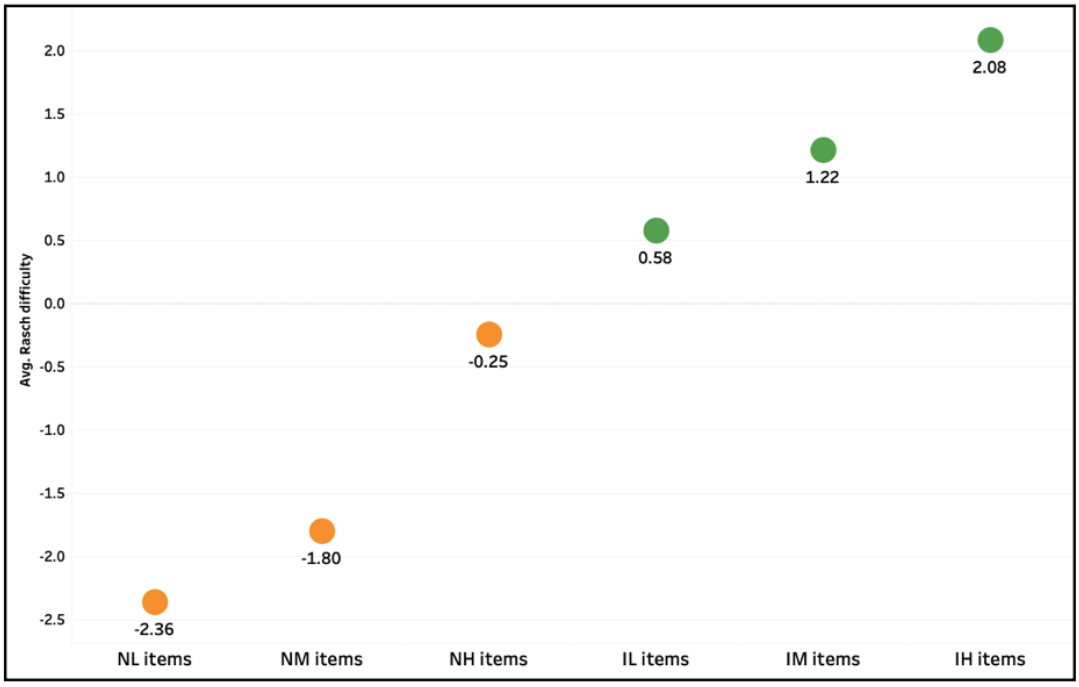

Figure 8. *Average Rasch difficulty of STAMP 4Se Chinese Reading items at levels 1 (Novice Low – NL) through 6 (Intermediate High – IH)*

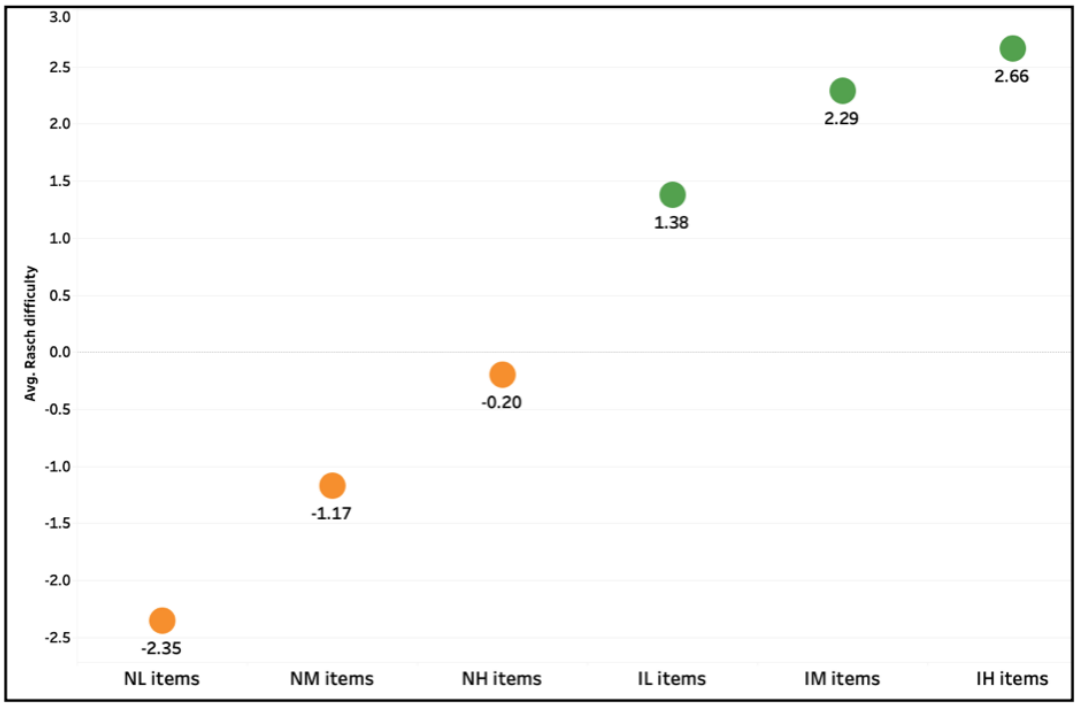

Figure 9. *Average Rasch difficulty of STAMP 4Se Chinese Listening items at levels 1 (Novice Low – NL) through 6 (Intermediate High – IH)*

STAMP 4Se French

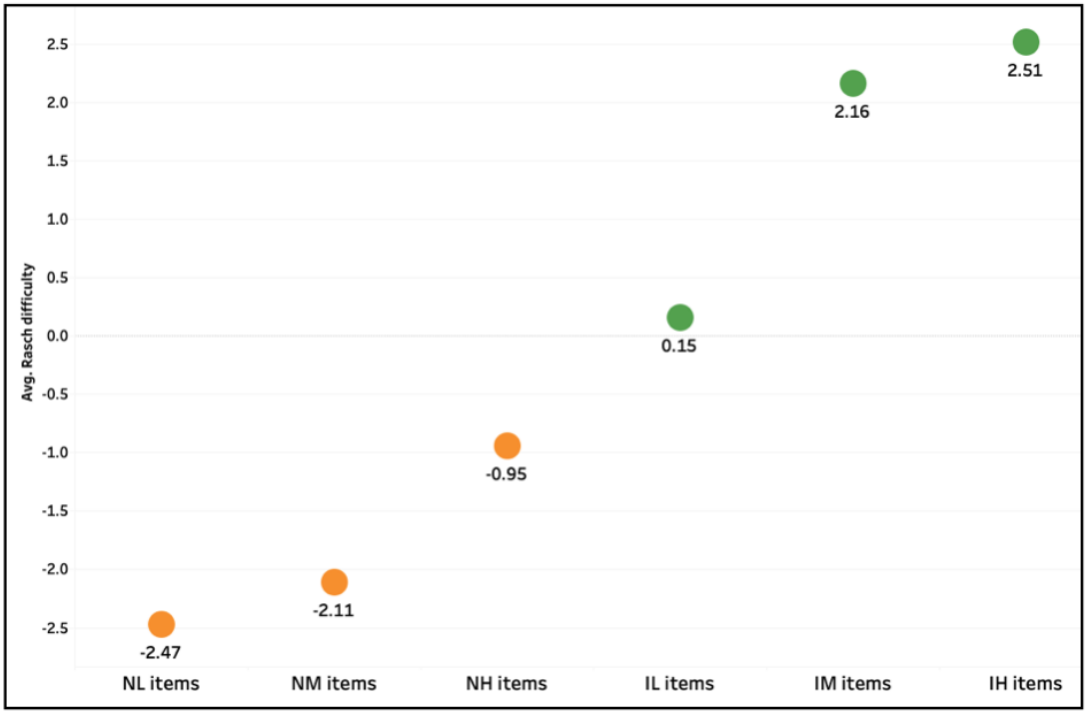

Figure 10. *Average Rasch Difficulty of STAMP 4Se French Reading Items at levels 1 (Novice Low – NL) through 6 (Intermediate High – IH)*

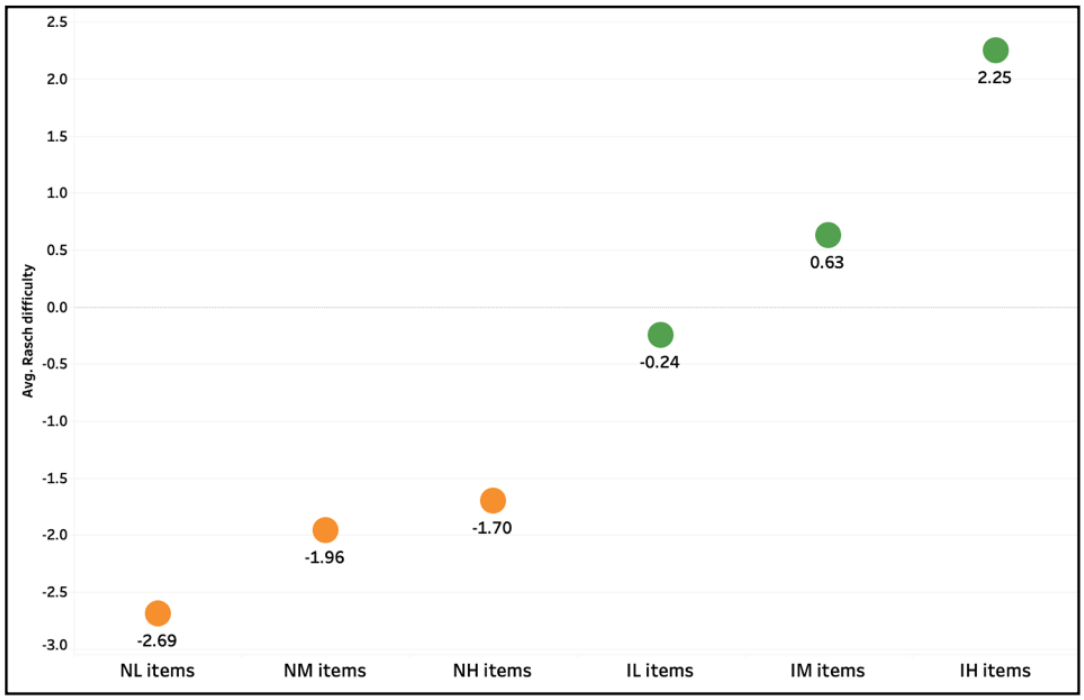

Figure 11. *Average Rasch difficulty of STAMP 4Se French Listening items at levels 1 (Novice Low – NL) through 6 (Intermediate High – IH)*

Rating Accuracy in the Writing and Speaking sections of STAMP 4Se and Inter-Rater Reliability

Teachers can log in and see their students' spoken and written responses that were rated based on the rubric used by certified Avant Assessment raters. The same rubric is used for rating all Speaking and Writing items. Writing and Speaking scores are determined by Avant-trained raters who 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 high Inter-Rater-Reliability (IRR) and that raters are rating accurately and not drifting from the standards, 20% of all responses are graded by a second rater. In the rare case where the two raters disagree with one another, the response is rated by a third rater who serves as a tiebreaker, so that an official score can be awarded to that response. Managers continuously monitor rating of all raters to ensure a high level of rating accuracy by Avant raters. 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.

It is important to understand that it is not viable to always expect perfect agreement between two human raters. Despite all the training they each may have gone through and all the experience and expertise each one may have regarding the construct being evaluated (in our case, language proficiency), even highly qualified humans disagree at times. Doctors do it. Engineers do it. Scientists do it. Therefore, the idea is to aim for as high an agreement as is feasible, and which proves defensible given the uses and interpretations of the scores from that test.

In addition to looking at the performance of individual raters, Avant also continuously monitors the quality of the rating at a group level. Below are the statistical measures that we at Avant Assessment run on the STAMP 4se test in order to assess the quality of the rating provided by our team of human raters as a group, and to assess whether the rating standards are high enough. While many companies may only report exact and adjacent agreement, we consider additional measures as well, since any specific measure can only provide partial information as to the quality of the raters. The more measures included, the more we can triangulate the results and arrive at a conclusive decision.

Exact Agreement: This measure is reported as a percentage that indicates the percentage of times, across the entire dataset analyzed, when the level awarded to a given response by Rater 1 is the same as the level awarded by Rater 2. For example, if Rater 1 awards a STAMP level 5 to a response and Rater 2 also awards a STAMP level 5 to that same response, that would be considered an instance of exact agreement. Feldt and Brennan (1989) suggest that when two raters are used, there should be an exact agreement of at least 80%, with 70% being considered acceptable for operational use.

Exact + Adjacent Agreement: This measure is reported as a percentage that indicates the percentage of times, across the entire dataset analyzed, when the level awarded to a given response by Rater 1 is either exact or adjacent to the level awarded by Rater 2. For example, a STAMP level 5 is adjacent to both a STAMP level 4 and a STAMP level 6. Therefore, if Rater 1 assigns a STAMP level 4 to a response and Rater 2 assigns a STAMP level 5 to that response, this would count towards this measure, since these two levels are adjacent to each other. Graham et al. (2012) suggests that when the rating scale has more than 5-7 rating levels, as is the case with the STAMP scale, exact + adjacent agreement should be close to 90%.

Quadratic weighted kappa (QWK): Cohen's kappa, or κ, measures reliability between two raters by considering the possibility of agreement occurring by chance. For example, since the numerical STAMP scale in Writing and Speaking is a 9-point scale, going from STAMP level 0 (No Proficiency at all) to STAMP level 8, there is a 11.11% chance that any two raters would perfectly agree on a score simply by chance. At Avant, in addition to taking this chance agreement into account, we use quadratic weights when calculating kappa, which means a higher penalty is assigned to scores that are farther away from each other. In other words, observing a difference between a STAMP level 3 and a STAMP level 7 between two ratings to the same response is more problematic than observing a difference between a STAMP level 3 and a STAMP level 4. Williamson et. al. (2012) recommends that QWK must be >= 0.70 and Fleiss (2003) notes that values above 0.75 show excellent agreement beyond chance for most purposes. A QWK value of 0 indicates agreement simply at the level of chance between two sets of ratings whereas a value of 1 indicates perfect agreement.

Standardized Mean Difference (SMD): This measure shows the extent to which two raters may be using a rating scale in a similar way. It shows the difference of the mean of two sets of scores (i.e., Rater 1 vs. Rater 2) standardized by the pooled standard deviation of those two sets. Ideally, neither rater should prefer or avoid awarding levels at a certain point of a rating scale (for example, avoid giving either STAMP 0s or STAMP 8s). In other words, both raters should make equal use of the rating scale and the scores awarded should be dependent only on the level of proficiency shown in the response itself. It is recommended that the value for this measure should be <= 0.15 (Williamson et al., 2012), ensuring that the distribution of both sets of scores is acceptably similar.

Spearman's Rank-Order Correlation (ρ): This measure indicates the strength of association between two variables, in this case the STAMP level assigned by Rater 1 and the STAMP level assigned by Rater 2. It is expected, if the team of raters are well trained and clearly understand the rating rubric, that whenever Rater 1 assigns a high proficiency level to a response, Rater 2 would also assign a high level. In other words, we expect the two sets of scores to move together (up or down) if the raters are indeed evaluating the same construct. We use Spearman's rank-order correlation coefficient instead of Pearson product-moment correlation since the former is preferred when the ratings are ordinal, as in the case of STAMP proficiency levels. A correlation coefficient of 0.80 or above is considered to be strong across various fields (Akoglu, 2018).

We now turn our attention to the quality of the ratings, in view of the statistics above, for the Writing and Speaking sections of STAMP 4Se across the same three representative languages: Spanish, French, and Chinese (Simplified). We provide below results based on two different sets of comparisons:

Rater 1 vs Rater 2: We compare the STAMP level awarded by Rater 1 to the STAMP level awarded by Rater 2 across a large number of responses in that language that were rated by at least two raters. This provides support for the reliability of the ratings provided by two randomly assigned Avant raters. As previously mentioned, two raters could award the exact same STAMP level to an essay and both could still be incorrect in their rating, vis-a-vis what the actual rating should have been for that response. For that reason, we do not include exact agreement measures between Rater 1 and Rater 2. Instead, we focus on Exact + Adjacent Agreement and also report on accuracy measures between the score awarded by Rater 1 (who rates solo 80% of the time) and official scores.

Rater 1 vs Official Score: To assess the accuracy of the levels assigned by Avant raters to responses, we look at a large number of instances in which a response was scored by two or more raters. We then compare the official score assigned to that response in the system (which is derived from the individual ratings for that response, as previously explained) to the score assigned by Rater 1 only. This provides us with an indication of how accurately a response is rated when only one Avant rater rates a response (which happens 80% of the time). Tables 2 and 3 show the statistical measures for the Writing and Speaking sections of STAMP 4Se Spanish, French, and Chinese (Simplified).

STAMP 4Se Language Writing	**Exact Agreement** *(Rater 1 vs. Official Score)*	**Exact + Adjacent Agreement** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Quadratic Weight Kappa (QWK)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Standardized Mean Difference (SMD)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Spearman's Rank-Order Correlation (ρ)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*
Spanish	*(88.6%)*	99.02% *(99.63%)*	0.96 *(0.98)*	0.006 *(-0.009)*	0.96 *(0.98)*
French	*(84.53%)*	99.04% *(99.65%)*	0.94 *(0.96)*	-0.02 *(0.02)*	0.95 *(0.96)*
Chinese Simplified	*(84.70%)*	97.69% *(99.51%)*	0.94 *(0.97)*	0.00 *(-0.01)*	0.94 *(0.97)*

Table 2. Inter-rater reliability statistics for STAMP 4Se Writing.

STAMP 4Se Language Speaking	**Exact Agreement** *(Rater 1 vs. Official Score)*	**Exact + Adjacent Agreement** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Quadratic Weight Kappa (QWK)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Standardized Mean Difference (SMD)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*	**Spearman's Rank-Order Correlation (ρ)** *Rater 1 vs. Rater 2 (Rater 1 vs. Official Score)*
Spanish	*(85.20%)*	96.51% *(98.24%)*	0.92 *(0.96)*	-0.00 *(-0.00)*	0.93 *(0.96)*
French	*(85.84%)*	98.06% *(98.88%)*	0.94 *(0.96)*	-0.03 *(0.02)*	0.94 *(0.96)*
Chinese Simplified	*(79.91%)*	96.23% *(98.50%)*	0.90 *(0.95)*	-0.00 *(-0.01)*	0.91 *(0.95)*

Table 3. Inter-rater reliability statistics for STAMP 4Se Speaking.

References

Akoglu, H. (2018). User's guide to correlation coefficients. Turkish journal of emergency medicine, 18(3), 91–93.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 105–146). New York: Macmillan.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions. 3rd ed. Wiley.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational measurement: issues and practice, 31(1), 2-13.

Appendix 1 – 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 2 – 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 3 – Table 1 STAMP 4Se Rubric

LEVEL	TEXT TYPE	LANGUAGE CONTROL		
		Functions/Complexity	*Vocabulary*	*Accuracy/Comprehensibility*
Novice-Low (STAMP Level 1)	**Words** Shows ability to produce individual words that could be related to the prompt.	Use of isolated words that deal with the prompt/task, shows inability to connect words in order to create meaning.	Limited vocabulary which deals with the prompt or situation.	Errors in spelling, word order, word choice and usage limit communication. Language produced can only be understood by the reader/listener with great effort by someone accustomed to a language learner
Novice-Mid (STAMP Level 2)	**Phrases** Shows ability to create simple meaning by grammatically connecting words. Specifically, some basic subjects and verbs or verbs and objects, but may be inconsistent at doing this.	Single, isolated connections to verbs. May be inconsistent at connecting words grammatically or have errors throughout. However, the errors must not prevent understanding of what is being said.	Typically limited in their vocabulary to Novice level topics that they experience in every-day life or that they have recently learned.	Errors in grammar, word order and word choice are prevalent and limit communication. Language produced is understood with difficulty by someone accustomed to a language learner.
Novice-High (STAMP Level 3)	**Simple Sentences** Shows ability to create simple sentences with very basic grammatical control.	Shows the ability to use very simple structures and functions of the language that have just been learned or studied. Extensive use of formulaic sentences, phrases and memorized sayings.	Generally, sentences that are created use basic vocabulary words with limited ability to elaborate.	Errors in grammar, usage, word order, and word choice sometimes limit communication. Language produced is mostly understood by someone accustomed to a language learner with some effort.

LEVEL	TEXT TYPE	LANGUAGE CONTROL		
		Functions/Complexity	*Vocabulary*	*Accuracy/Comprehensibility*
Intermediate-Low (STAMP Level 4)	**Strings of Sentences** Shows ability to create simple sentences with some added detail. Simple sentences with different forms of added detail are generally produced with no connections or links to each other.	Shows the ability to produce simple sentences that are enhanced through the use of prepositional phrases, adverbs, etc. Independent sentences (ideas) can be moved around without affecting the overall meaning of the response.	Vocabulary is beginning to expand beyond the most frequent words and the ability to elaborate is more evident in the language produced. Drawn from daily life.	Errors in usage, grammar, word order, and word choice continue to be common, but generally do not hinder communication. Language produced is understood by someone accustomed to a language learner with little effort.
Intermediate-Mid (STAMP Level 5)	**Connected Sentences** Shows ability to create enough language to address a majority of the prompt or situation, showing groupings of ideas. Thoughts are loosely connected and generally cannot be moved around without affecting meaning.	Demonstrates the ability to create enough language that shows the beginning of connectedness. Able to create several sentences with complexity and may use some transition words. Connectedness begins to emerge as they create 'groupings of sentences.' Learners begin to transfer previously learned skills and language to new structures/functions.	Vocabulary use is expanding, and language used is more than just the usual, high frequency or most commonly used vocabulary. May begin to use circumlocution haltingly due to limited vocabulary.	Shows ability to use more than just simple present tense, however errors occur when trying to use other tenses. New skills, such as creating more complex sentence structures or using other tenses, will generate some errors. Language produced is easily understood by someone accustomed to a language learner.
	Pre-Paragraph Shows ability to create language with	Shows the ability to use different time frames and just beginning to develop the ability to	Use of transition words and concepts with more ease is evident in language production. Circumlocution	At this level, good control of the language and confidence is evident with an increasing range of topics. There are still occasional errors in language production, but errors do not hinder

Intermediate-High (STAMP Level 6)	a more natural flow. The increased number of complex structures are well constructed. Sentences and ideas are connected with multiple, varied connectors, transitions and other linking strategies.	switch most time frames (present, past and future) with increased accuracy. Complexity and variety of sentence types and structures is increasing, helping move response to a more natural and smooth flow.	is used more effectively. Ability to create new language on less common topics is evident.	ability to communicate. Language produced is generally understood by someone accustomed and those unaccustomed to a language learner.
LEVEL	**TEXT TYPE**	**LANGUAGE CONTROL**		
		Functions/Complexity	***Vocabulary***	***Accuracy/Comprehensibility***
Advanced-Low (STAMP Level 7)	**Paragraph/Advanced Language** Shows ability to address each aspect of the response with Complex structures, which demonstrate an increasing ability to produce a greater depth of meaning with language that effectively and more thoroughly addresses each aspect of the prompt. Able to create a paragraph-length description with a natural flow.	Shows the ability to create a smooth and natural flow by using a variety of added details, complex grammar and descriptive language. Shows ability to switch time frames naturally with a high degree of accuracy. Ability to use a wide variety of sentence structures, patterns and tenses is evident in communications.	Use of advanced vocabulary (less frequent and specialized), advanced structures and/or terms evident. Able to address a wide variety of 'less common' topics. Advanced language is used within the response, which helps demonstrate an increased ability to demonstrate their language skills more effectively.	Majority of language is error-free, creating a smooth and natural flow. However, there may still be occasional errors, but without pattern or causing any breakdown in communication. Language produced is easily understood by those unaccustomed to language learners. Use of correct orthography (elements of writing such as spelling, grammar, punctuation, accents, tonal markers, umlauts etc.) increases in importance – especially if the desire is to reach Advanced levels. Correct orthography is expected to meet basic WORK and/or academic writing needs at the Advanced level.
	Extended Paragraph and Language	Shows the ability to create sophisticated	Effective use of concise language across a wide	Language is almost entirely error free, creating a smooth and natural flow. Any errors in the

Advanced-Mid (STAMP Level 8)	Shows ability to confidently address each aspect of the prompt with clear organization and a native-like flow. Able to incorporate a significant number of complexities with higher degree of accuracy throughout, giving that depth of meaning expected at Advanced Mid. Shows skill with creating a response that is interwoven with lexical and syntactic density, which one might expect to see at the Advanced level. Increasing ability to extend discourse beyond immediate experience to better address the prompt.	language with in-depth description and narration interwoven throughout. Syntactic density is evident as well. Ability to switch time frames is natural and generally without error. Complex structures and grammar are used to create linguistic diversity in the language.	variety of topics is evident. Vocabulary helps create Advanced language throughout the response, demonstrating a deeper cultural understanding, adding more clarity and depth of meaning.	language are not easily identified and do not occur in any patterned way. Language produced is native-speaker-like and understood by those unaccustomed to language learners. Uses correct orthography throughout the response. Correct orthography is expected to meet basic WORK and/or academic writing needs at the Advanced level.