



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

**Arabic 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

6/6/2012

*\*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 *Arabic Computerized Assessment of Proficiency (Arabic Avant STAMP 4S)* published by CASLS (Technical Report 2010-3). 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 Arabic. The development of this test was funded through the National Security Education Program (NSEP) with the mandate of developing an assessment in Modern Standard Arabic in reading, listening, writing and speaking based on the existing infrastructure for the Standards-based Measurement of Proficiency (STAMP), a previous CASLS project to develop online proficiency tests in modern foreign languages. Avant STAMP 4S is an online proficiency-oriented test of listening, reading, writing and speaking.

This document has several major sections. The first and second sections give an overview of Avant STAMP 4S and format of the test. The third section details the development of the test items. The fourth describes the technical characteristics of the final test. The fifth section presents information on how the test is scored.

### **Acknowledgment**

The development of this test was made possible 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 .....	4
Preface .....	5
Executive Summary .....	6
1 Overview and purpose of the assessment .....	7
1.1 Construct for STAMP 4S.....	7
1.2 Test level .....	7
1.3 Population served by the assessment .....	10
2 Description of the assessment.....	10
2.1 Content and structure of the STAMP 4S.....	10
2.2 Test delivery.....	11
3 Test development .....	12
3.1 Item writing.....	13
3.2 Internal review and revisions .....	14
3.3 Graphics development.....	14
4 Technical characteristics.....	16
4.1 Field testing.....	16
4.2 Selection of items .....	17
4.3 Preparation for delivery.....	17
4.4 Determination of cut scores .....	18
5 Score reporting .....	18
5.1 Reading and listening scores .....	20
5.2 Writing and speaking scores.....	20
References .....	23
A Rasch summary results – reading.....	24
B Rasch summary results – listening .....	27
C Bin information .....	28

## Nomenclature

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

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 Arabic, an online assessment of Modern Standard Arabic 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) 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 speakers of Arabic identified reading and listening passages and CASLS staff wrote corresponding items. A comprehensive review of the test items was conducted in June 2008. Reviewers expressed general satisfaction with the test items, though there were discrepancies between the intended proficiency level and the reviewers' estimation of the level. This was most evident with items geared towards the upper proficiency levels. The most promising items were selected for field testing.

Empirical information on the items was collected through an adaptive field test. Approximately 1500 students participated in field testing. Speech and writing samples were collected for those test sections, but no ratings were given. Reading and listening data from the field tests was analyzed using a Rasch methodology. The person reliability from the reading item analysis was estimated at .93 and the person reliability from the listening item analysis was .89. Appropriately functioning items were assembled into a test panel using empirical information to establish a score table and routing table. Cut scores for proficiency levels were set according to the external review. Simulations of the delivery algorithm show a correlation of  $r = .98$  between simulated test-taker ability and final ability estimate on the operational panel. The simulation also shows that the reading section is 90% accurate and the listening section is 89% accurate in identifying the students' "true" proficiency levels.

# 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 based on expert review can be found in Section 5.1.

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
	Level 9	2+	Advanced-High
Expanding	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).

Arabic 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 Arabic Avant STAMP 4S.



**Table 2**

Language Proficiency Measured by Avant STAMP 4S (based on Bachman &amp; 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 Standard Arabic, 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 Arabic 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

Arabic Avant STAMP 4S is designed to provide a general overall estimate of a language learner's proficiency in four skills in Arabic. 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. 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 it is 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 Arabic 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 Arabic 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 Arabic text as well as contextualizing information. The test question, in English, requires the examinee to read the information in Arabic and choose the best answer from the options provided. Examinees must answer the question before proceeding to the next screen. Backtracking is not allowed.

**Situation**

You are traveling near the Red Sea when you see this sign:



**Question 1 / 1**

What does this sign prohibit?

- fishing
- swimming
- boating
- camping

*Figure 1. Arabic reading item*

Arabic 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 Arabic 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 Arabic 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.

**Situation**

Members of your class are introducing themselves.



**Question 1/1**

How many languages does Layla mention?

CLICK PLAY WHEN READY

four  
 three  
 two  
 one

Figure 2. Arabic listening item

A visual depiction of the Arabic Avant STAMP 4S algorithm is shown in Figure 6 on page 24.

### 3 Test development

The general test development process for Arabic Avant STAMP 4S is illustrated in Figure 3.

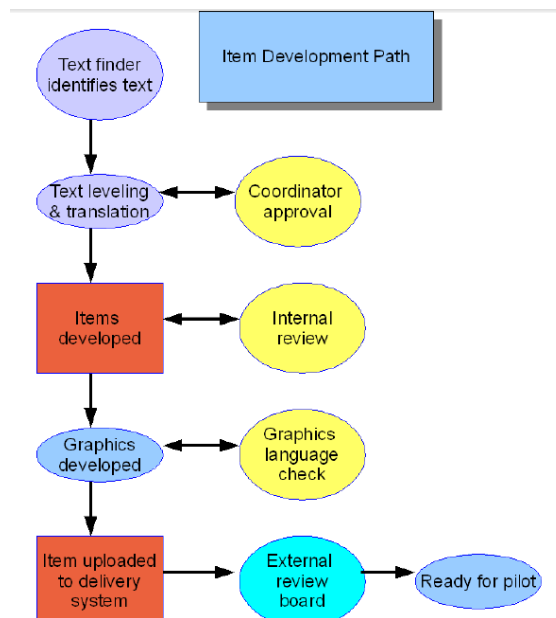


Figure 3. Item writing workflow

### 3.1 Item writing

General test specifications for Avant STAMP 4S were developed based on the Avant STAMP 4S framework documents. Additional input in the form of Arabic-specific test specifications were provided by the Center for Applied Linguistics (CAL). Task specifications were then created to further elucidate the contents of the assessment. The complete test and task specifications are available on the CASLS website.<sup>1</sup> Existing item writing guides and job aids were updated as necessary.

CASLS hired four native Arabic-speaking students to initially develop content for this project and serve as “text finders.” Prior to beginning work, all CASLS’ staff involved in the project was trained to rate texts according to ILR levels using the self-study *Passage Rating Course* designed by the National Foreign Language Center (NFLC). This training was supplemented with meetings to discuss the levels of texts that had been created or adapted from authentic texts. The Arabic-speaking students came from geographically diverse Arabic-speaking areas: the Gulf, North Africa and Belad Al-Sham.

For lower-level items, text finders created reading and listening texts that best matched the test specifications and target proficiency levels. Especially in the case of listening, this involved developing original material, as most “everyday” spoken interactions would be performed in the regional dialect and not Modern Standard Arabic. Draft passages deemed worthy or further development were uploaded into an internal item bank database.

For advanced level texts, text finders were tasked with finding authentic reading and listening texts that best matched the test specifications and target proficiency levels. This was primarily done by searching through Arabic-language resources on the World Wide Web. Many authentic texts could be discounted out of hand as being too long or requiring too much background information. Texts that seemed promising were saved for translation. In the case of audio texts, this usually required identifying portions of longer audio files. Though the text finders scoured many websites for texts, only a small portion of texts reviewed were kept and translated. Of those “found” texts, only a subset was considered good enough to be used in item development. Table 3 shows the number of texts used.

Table 3  
Advanced Arabic Text Counts

Skill	Texts Found	Texts Developed	Items Created
Reading	66	20	75
Listening	37	19	59

Finding appropriate Refining (ACTFL Superior / ILR 3) texts proved especially challenging. For this reason, effort was concentrated on the levels up to Expanding (ACTFL Advanced / ILR 2). Table 4 shows the number of items developed for the reading and listening sections.

A set of four speaking and writing prompts was created by CASLS staff. As the speaking and writing prompts are delivered in English, CASLS uses similar prompts across languages. These prompts were written to allow for responses across the range of proficiency levels.

<sup>1</sup> <http://casls.uoregon.edu>

### 3.2 Internal review and revisions

Throughout the item development process, items were subject to internal review with text finders acting as reviewers. For each Arabic text in the item database, reviewers completed the checklist shown in Table 5. Text finders were instructed to complete the checklist for texts that had been found by other text finders. Items or passages that could not satisfy all of the criteria were marked with a “Problem found” tag for further review by the test developer. A comment box was available for each item to allow reviewers to make annotations about specific problems.

Table 4  
Item Counts for Reading and Listening

Skill	Level	Passages	Items
Reading	Refining	7	20
	Expanding	13	51
	Transitioning	26	34
	Beginning	58	60
Listening	Refining	3	4
	Expanding	16	55
	Transitioning	27	58
	Beginning	59	79

Table 5  
Text and Item Review

Criteria	Checked
Text-specific	
The text is authentic	<input type="checkbox"/>
The text/task fits the life experience of the target audience	<input type="checkbox"/>
The level of the text/task is appropriate	<input type="checkbox"/>
The audio quality is acceptable (listening)	<input type="checkbox"/>
Item-specific	
The question focuses on important and not trivial content	<input type="checkbox"/>
There is only one correct answer	<input type="checkbox"/>
All distractors are plausible	<input type="checkbox"/>
The question is independent of others	<input type="checkbox"/>
The question matches the level of the passage	<input type="checkbox"/>
The item can be answered without specialized background knowledge	<input type="checkbox"/>
The item is free from bias	<input type="checkbox"/>
The task is open ended (writing/speaking)	<input type="checkbox"/>

### 3.3 Graphics development

Because the test is intended to be compatible with any computer, CASLS renders Arabic 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 Arabic-speaking text finders reviewed finished items to ensure that the text was being correctly displayed in the final item. The left-to-right nature of the Arabic text sometimes created formatting difficulties when transferring text, requiring several rounds of discussion between the item reviewer and graphic artists for some of the images.

### **3.4 Revisions**

After an external review (see Section 5.1), CASLS staff corrected the Arabic texts and items that had been classified as problematic during the review. Even though many of the texts were based on authentic sources, the reviewers found some spelling or grammar errors to be corrected. Corrections to the texts involved revising the Arabic text in a word processing program, taking a screenshot, and remaking the graphic to incorporate the new text. Some audio passages were also rerecorded to correct infelicities in the originals.

A total of 111 reading items and 163 listening items were developed and uploaded into the Avant STAMP 4S testing system as a result of this item development process. Four speaking and four writing prompts were also uploaded to Arabic Avant STAMP 4S. Additional speaking and writing prompts have been added by Avant Assessment to the initial pool.

## 4 Technical characteristics

### 4.1 Field testing

Field testing was conducted over a multiyear period beginning in October 2007. This long field testing window was needed to accommodate the realities of the academic schedule and give participant sites maximum flexibility in choosing pilot test dates. The data analyzed for this report was collected through June 1, 2010.

#### Participants

CASLS did not solicit specific schools to participate in field testing, but rather allowed any willing program to register for the test. No biodata was collected from individual students, though it is assumed that those programs registering for the field test would be those programs with an interest in the finished test as well. Approximately 1500 students participated in field testing<sup>2</sup>. Figure 4 shows a map of the relative number of field test participants by state.

#### Materials

A set of 80 reading and 91 listening items were chosen to field test. These items were chosen for having “passed” the review with no or minor revisions, thus best representing the intended construct and proficiency levels. These items were arranged into bins of between 8 - 15 items across three levels of relative difficulty in a “floor first” adaptive design (Figure 5). Since difficulty estimations were not available for these items, routing tables were created using percentage correct at level rather than item information intersections. A score table with tentative proficiency estimates was also created using percentage correct at level. These scores were provided as a service to teachers to provide tentative feedback about their students. Four speaking and four writing prompts were also made available for the field test.

#### Results

Test results were analyzed with the Rasch analysis program, Winsteps (Linacre, 2008). Summary data is presented in Appendix C. The person reliability estimate was .93 for reading and .89 for listening. The person separation values of 3.54 and 2.84 for reading and listening respectively indicate that the test is not sensitive enough to distinguish the nine levels of proficiency that would correspond to low, mid and high sublevels<sup>3</sup>. For this reason, only major proficiency level distinctions are reported. Results of this analysis were used to estimate the item difficulties for the final routing and scoring tables. For the final calibration run, one misfitting item and nine misfitting responses were eliminated from the reading data; for listening, one misfitting item and three misfitting responses were omitted.

<sup>2</sup> Because CASLS’ system adheres to human subjects’ protections by tracking only test instances and not individuals and many participants may have taken multiple skills, it is impossible to determine exactly how many individual students participated. It is assumed that many students took multiple skills and that some students may have taken the same skill more than once over the course of field testing. This number is a conservative estimate based on the number of tests delivered and assuming some overlap between skills.

<sup>3</sup> From the Rasch separation value it is possible to compute the number of *strata*, or statistically distinct level of performance using the formula  $H = (4G+1) \div 3$ , where G is the separation index.



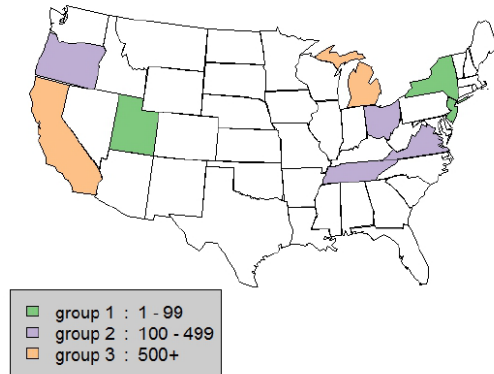


Figure 4. Map of Arabic field test participants

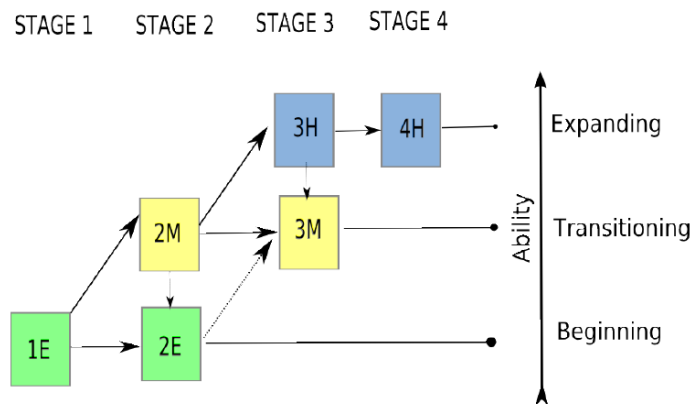


Figure 5. "Floor first" delivery

## 4.2 Selection of items

Not all of the items developed for the test have been included in the operational form. Items that passed internal reviews were used in field testing. Rasch analysis of those test results produced difficulty estimates for each of the items. Items with mean squared infit values between .5 and 1.5 were considered acceptable for inclusion in the pool. In some cases, this meant that not all items in an item set<sup>4</sup> were included in the operational pool. The difficulty values of these items will be used as anchor values when calibrating new items into the pool in the future.

## 4.3 Preparation for delivery

An iterative process was used to place listening and reading items in bins for multistage delivery. The goal was to create bins of 10 items each. The multistage delivery paradigm involves routing the test taker through bins of varying relative difficulty based on which bin will provide the most information about the test taker's ability given their performance on the previous bin.<sup>5</sup> Thus, a test taker who has answered many questions successfully in a given bin will get a more challenging bin in the next stage; a test taker who has not answered many questions successfully will get a bin at a similar or easier level in

<sup>4</sup> A common passage with more than one associated question.

<sup>5</sup> For Rasch-based tests, the most informative item is one for which the test taker has a 50% probability of success.

the next stage. (See Figure 6 for a graphical representation.) However, because many items were part of an item set it was not always possible to create the optimum arrangement to maximize bin information, as items in an item set cannot be split across bins.<sup>6</sup>

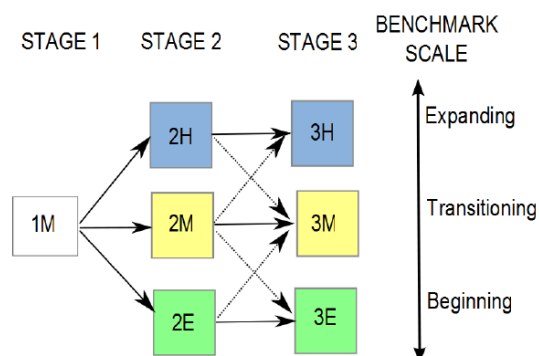


Figure 6. Delivery algorithm

#### 4.4 Determination of cut scores

For listening and reading sections, cut scores were determined by calculating the median item difficulty for the items assigned to each major proficiency level from those items remaining in the pool. A value of 1:4 logits was added to this value to determine the ability level needed to have an 80% probability of answering a median level question correctly. The exception to this is the cut score for Beginning, which is set to just above chance. Maintaining cut scores as Rasch values rather than using number of items correct allows the particular items in the test to change while the cut scores stay stable.

## 5 Validity evidence

### 5.1 External review

A comprehensive review for the Arabic Avant STAMP 4S was held at CASLS on June 7 - 8, 2008. The purpose of the review was twofold:

- 1) To have the quality of the items reviewed by independent experts, and
- 2) To provide evidence that the items were appropriate for the proficiency levels targeted.

#### Participants

The following Arabic specialists participated as external reviewers in the two-day on-site session:

- Dr. Mahdi Alish (United States Military Academy)
- Dr. Salah Ayari (Texas A & M University)
- Dr. Hanada Taha-Thomure (San Diego State University)

Two additional specialists provided external reviewing separately using CASLS online system.

- Dr. Naji A. Jabar (Bridge Academy)
- Dr. Wafa Hassan (Michigan State University)

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

<sup>6</sup> An example of bin information functions can be seen in Appendix D.

## Procedure

Onsite reviewers were given an overview presentation of the background and design of the test and allowed to ask any clarification questions they had. They were then asked to log on to the test system on individual workstations and mark their perceived proficiency level of each item on a sheet corresponding to the item. It was felt that the use of hard copies to record ratings would be more efficient for note-taking and commenting. (See sample in Appendix B.)

The items were delivered in a set of five rounds over two days. Although the review plan originally called for completing the review in one day and doing a modified Angoff rating session on day two, it quickly became apparent that the reviewers would need the full two days to review all the items. Each round consisted of approximately 40 items, split between listening and reading. The items were identified only by an ID number and no indication of their intended proficiency level was provided. Each reviewer saw the same items, but the order of presentation was randomized automatically by the delivery system. Items targeting a range of proficiency levels were presented in each round. After each round, participants reconvened as a group to discuss problematic items and compare ratings. The Arabic-speaking text finders were present during the review session to take notes on any changes to be made to the Arabic text.

Table 6  
Correlations between proficiency ratings

	CASLS	Rater 1	Rater 2	Rater 3
CASLS	1.00			
Rater 1	0.81	1.00		
Rater 2	0.88	0.78	1.00	
Rater 3	0.85	0.78	0.82	1.00

## Results

Overall, the reviewers expressed general satisfaction with the test design and items. There was also a fairly high level of agreement between CASLS targeted item level and the reviewers' ratings as shown in Table 6.

The reviewers did express some areas for improvement, however. Although all of the authentic texts used for the items were written in Arabic and intended for an Arabic-speaking audience, several of them were apparently translations of texts originally written in other languages. As a result, some of the texts had grammatical problems and non-standard discourse styles resulting from sloppy translation. The reviewers suggested that these texts be corrected or eliminated. There reviewers also expressed concern with the quality of presentation of some of the listening passages and suggested that they be re-recorded. The reviewers also noted that many of the texts dealt with general issues and suggested more emphasis on Arab-related topics if possible.

## 6 Score reporting

Arabic 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.

## 6.1 Reading and listening scores

Reading scores and listening scores are reported as general proficiency levels and as scaled scores. For each possible path and total score combination on the test, a corresponding Rasch ability estimate is available in a score lookup table that the reporting system uses to generate final scores. This ability estimate is higher for examinees who have been routed to “harder” items on the test, even though their total raw score may be the same as other students seeing only less challenging items. The scaled score is derived by multiplying the Rasch ability 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 original logit scale values into a more user-friendly format and should be interpreted only in relation to cut scores for this test and not similar scores for other standardized tests. Cut scores are shown in Table 7.

**Table 7**  
Cut Scores for Scaled Scores

Level	Reading	Listening
Beginning	345	372
Transitioning	546	539
Expanding	655	657

There is approximately a  $\pm 22$  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.

## 6.2 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 8**  
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 9**

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.

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.

## A Standard-setting outline

### Arabic Standard-setting Agenda

#### Day 1 (Saturday)

Agenda Item	Goal	Comments	Time
Pick up at Hilton – walk to CASLS		Meet in Lobby	8:30
Introductions	Introduce participants	Take care of any outstanding paperwork needs	9:00
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 Avant STAMP 2.0 project; ACTFL/ILR guidelines, MSA issue	9:20
Questions / clarifications	Make sure any initial concerns/questions are addressed		10:00
Standard setting intro	Give participants overview of process; show procedure for Avant STAMP 4S login		10:10
Round 1	Raise participants' awareness of text versus item difficulty; make sure participants are "on the same page"	Working individually at computers	10:25
Round 1 Discussion	Participants bring up any issues that have arisen during the round	participants sort individually at first, then compare results	11:00
Round 2	Get consensus on results		11:45
Round 2 Discussion	Only discuss problematic items		12:15
Break for lunch	Discuss issues/questions in general term		12:30
Round 3	Participants work individually	Re-introduce ACTFL / ILR skill level descriptors	1:15
Round 3 Discussion	Only discuss problematic items		1:40
Round 4	Participants work individually		2:00
Round 4 Discussion	Only discuss problematic items		2:30
Break	Afternoon break		3:00
Round 5	Participants work individually		4:00
Round 5 Discussion	Only discuss problematic items		4:30
Speaking and Writing prompt review	Brief discussion of speaking / listening prompts		4:45
Wrap-up			5:30



## B External Item review

ARB-216-A

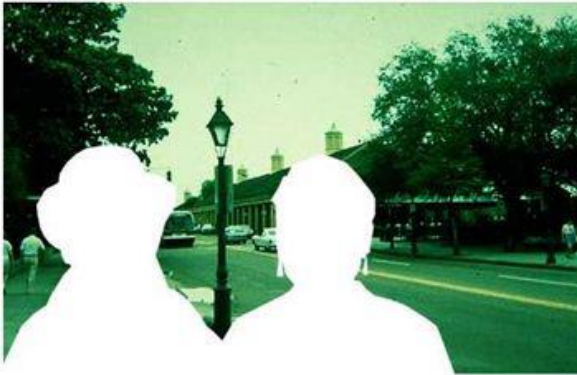
N		I		A		S		X



The Center for Applied Second Language Studies  
The Northwest National Foreign Language Resource Center

### Situation ARB-216-A

Amir and his friend Samira met in Egypt.



### Question 1/2

Why did Samira go to Egypt?



- to spend summer vacation
- to watch a song festival
- to visit family
- to buy clothes

[Close this window](#)

5290 University of Oregon, Eugene OR 97403-5290 Tel:541-346-5699 [info@darkwing.uoregon.edu](mailto:info@darkwing.uoregon.edu)

Copyright © 2008 Center for Applied Second Language Studies (CASLS)



Figure B.1. Rating sheet for external review.

## C Rasch summary results

Table C.1  
Arabic Reading Results - Persons

Summary of 1462 Measured (Non-Extreme) Persons

	Raw		Measure	Model	Infit		Outfit	
	Score	Count		Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	17.6	28.2	-.99	.49	1.00	.0	1.02	.1
S.D.	8.0	9.3	1.91	.12	.19	.9	.41	.9
Max	33.0	51.0	5.34	1.89	1.84	3.9	3.63	3.6
Min	1.0	2.0	-5.58	.35	.20	-2.5	.14	-1.9

*Note.* Winsteps v3.69 Table 3.1., Real RMSE=.52, TrueSD=1.84, Separation=3.54, Person Reliability=.93, Model RMSE=.50, TrueSD=1.85, Separation=3.68, Person Reliability=.93

Table C.2  
Arabic Reading Results - Items

Summary of 79 Measured (Non-Extreme) Items

	Raw		Measure	Model	Infit		Outfit	
	Score	Count		Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	327.5	526.8	.00	.17	.97	-.3	.96	-.3
S.D.	300.1	420.3	2.35	.13	.13	2.6	.23	2.4
Max	1007.0	1413.0	4.68	.73	1.41	9.9	1.79	9.6
Min	15.0	34.0	-4.52	.07	.67	-7.5	.47	-6.5

*Note.* Winsteps v3.69 Table 3.1., Real RMSE=.22, TrueSD=2.34, Separation=10.80, Item Reliability=.99, Model RMSE=.21, TrueSD=2.34, Separation=10.91, Item Reliability=.99

Table C.3  
Arabic Listening Results - Persons

Summary of 993 Measured (Non-Extreme) Persons

	Raw		Model Measure	Model Error	Infit		Outfit	
	Score	Count			MNSQ	ZSTD	MNSQ	ZSTD
Mean	18.8	30.8	-.58	.48	1.01	.0	.99	.0
S.D.	9.0	11.3	1.61	.17	.22	.9	.39	.9
Max	40.0	59.0	5.68	1.54	3.10	4.4	3.91	4.6
Min	1.0	2.0	-5.76	.33	.31	-2.9	.22	-2.2

*Note.* Winsteps v3.69 Table 3.1., Real RMSE=.53, TrueSD=1.52, Separation=2.84, Person Reliability=.89, Model RMSE=.51, TrueSD=1.52, Separation=3.00, Person Reliability=.90

Table C.4  
Arabic Listening Results - Items

Summary of 79 Measured (Non-Extreme) Items

	Raw		Model Measure	Model Error	Infit		Outfit	
	Score	Count			MNSQ	ZSTD	MNSQ	ZSTD
Mean	208.8	342.6	.00	.19	.98	-.2	.94	-.3
S.D.	176.9	283.4	1.92	.11	.12	2.2	.20	2.1
Max	694.0	979.0	4.49	.62	1.45	9.9	1.60	7.8
Min	10.0	25.0	-4.54	.08	.79	-5.1	.60	-4.5

*Note.* Winsteps v3.69 Table 3.1., Real RMSE=.22, TrueSD=1.90, Separation=8.78, Item Reliability=.99, Model RMSE=.21, TrueSD=1.90, Separation=8.86, Item Reliability=.99

## D Bin information

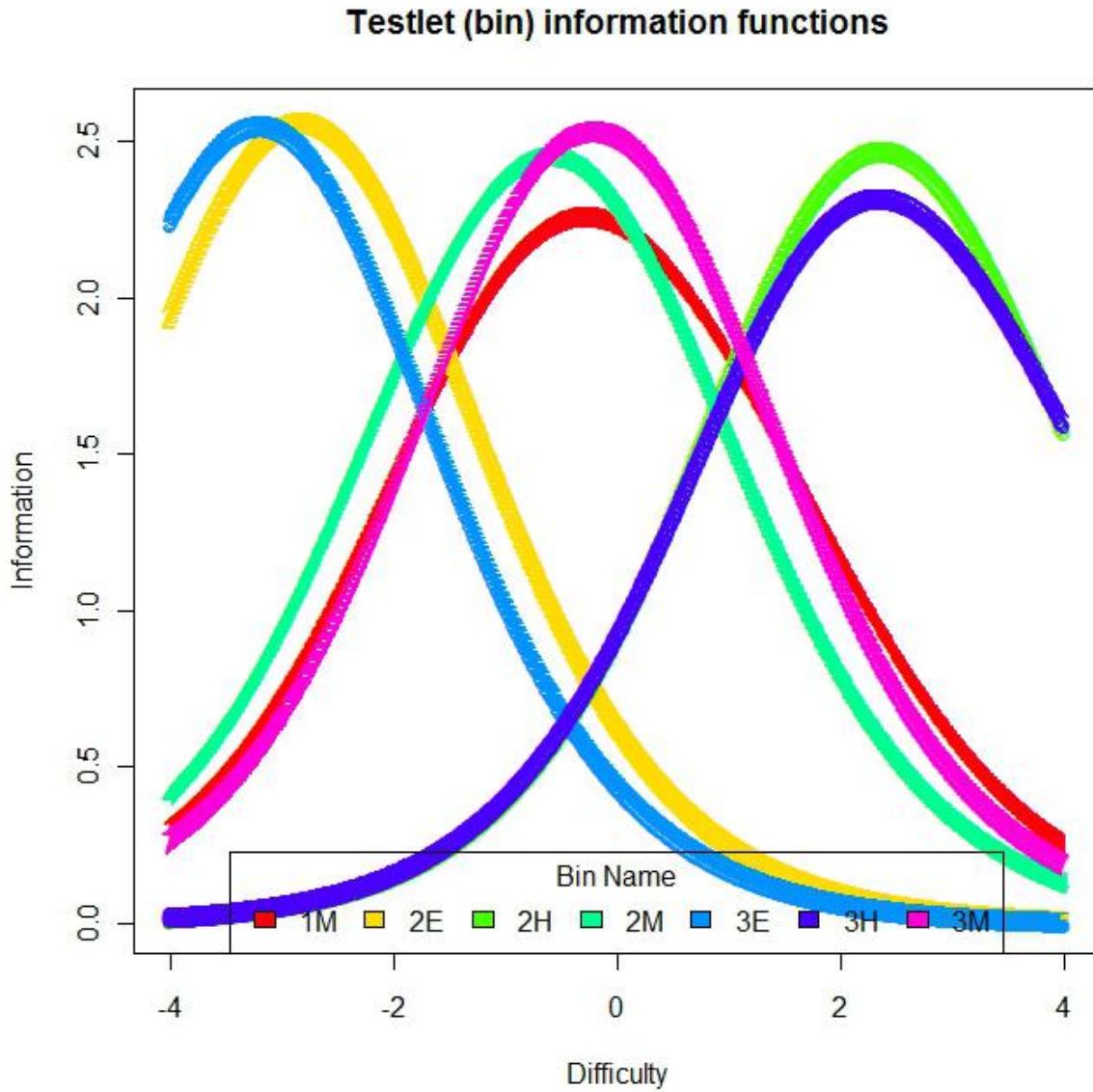


Figure D.1. Information functions intersections for reading test