CASLS REPORT

Technical Report 2010-2
Unlimited Release
Printed July 2010

Supersedes Persian Final Report
Dated NA

Persian Computerized Assessment of Proficiency (CAP)

Martyn Clark
Assessment Director

Prepared by
Center for Applied Second Language Studies
University of Oregon

CASLS, a National Foreign Language Resource Center and home of the Oregon Chinese Flagship Program, is dedicated to improving language teaching and learning.

Prepared by the Center for Applied Second Language Studies (CASLS).

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.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available from CASLS:

Campus:	5290 University of Oregon, Eugene OR 97403
Physical:	975 High St Suite 100, Eugene, OR 97401
Telephone:	(541) 346-5699
Fax:	(541) 346-6303
E-Mail:	info@uoregon.edu
Download:	http://casls.uoregon.edu/papers.php

Technical Report 2010-2
Unlimited Release
Printed July 2010

Supersedes Persian Final Report
Dated NA

Persian Computerized Assessment of Proficiency (CAP)

Martyn Clark
Assessment Director
martyn@uoregon.edu

Abstract

This document was prepared by the Center for Applied Second Language Studies (CASLS). It describes the development of the Persian Computerized Assessment of Proficiency (CAP). The development of this test was funded through the Kevorkian Center at New York University with the mandate of developing an assessment in Persian reading, 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.

This document has several major sections. The first and second sections give an overview of the Persian CAP project 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 validity evidence from an external review, pilot testing, and field testing.

Acknowledgment

The contents of this report 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 7
Preface 8
Executive summary 9
1 Overview and purpose of the assessment 11
1.1 Construct for the CAP 11
1.2 Test level 11
1.3 Population served by the assessment 14
2 Description of the assessment 15
2.1 Content and structure of the CAP 15
2.2 Test Delivery 16
3 Test development 17
3.1 Item writing 17
3.2 Internal review 18
3.3 Graphics development 18
3.4 Revisions 18
4 Technical characteristics 19
4.1 Field testing 19
4.2 Selection of items 20
4.3 Preparation for delivery 21
4.4 Determination of cut scores 21
5 Validity evidence 23
5.1 External review 23
5.2 Pilot Testing 25
6 Score reporting 27
6.1 Reading scores 27
6.2 Writing and speaking scores 27
References 29

Appendix

A Standard setting outline 30
B Rasch summary results 31

List of Figures

1 Persian Reading Item 16
2 Item Writing Workflow 17
3 Map of Persian field test participants 19
4 ”Floor First” Delivery 20
5 Delivery algorithm 21
6 Persian Pilot Results 26

List of Tables

1 CASLS Benchmark Levels 12
2 Language Proficiency Measured by CAP (based on Bachman & Palmer (1996)) . . . 13
3 Angoff Rating Correlations 24
4 Cut Scores for Scaled Scores 27
5 Common Speaking Rubric 28
6 Speaking Scores and Proficiency Levels 28

Nomenclature

ACTFL American Council on the Teaching of Foreign Languages

Avant Avant Assessment (formerly Language Learning Solutions)

Bin A group of test items delivered together

CAP Computerized Assessment of Proficiency

CASLS Center for Applied Second Language Studies

FSI/ILR Foreign Service Institute/Interagency Language Roundtable

Item set Two or more items sharing a common stimulus (e.g., a reading text)

LRC Language Resource Center

Level Level on a proficiency scale (e.g., Advanced-Mid)

Panel A term used to describe a particular arrangement of bins

Rasch A mathematical model of the probability of a correct response which takes person ability and item difficulty into account

Routing table A lookup table used by the test engine to choose the next most appropriate bin for a student

Score table A lookup table used by the scoring engine to determine an examinee's score based on their test path

STAMP *STA*ndards-based *M*easurement of *P*roficiency

Test path A record of the particular items that an examinee encounters during the test

Preface

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

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

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

The Computerized Assessment of Proficiency (CAP) 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 *STAMP* (*STA*ndards-based *M*easurement of *P*roficiency). We refer to tests under development as CAP to differentiate between research done by CASLS during the development phase from any additional work in the future by Avant Assessment.

Executive summary

CASLS has developed the Persian Computerized Assessment of Proficiency (Persian CAP), an online assessment of modern Persian that covers a proficiency range comparable to the American Council on the Teaching of Foreign Languages (ACTFL) proficiency levels Novice through Intermediate in three skills (reading, writing, presentational speaking). This test builds on the style and format of Standards-based Measurement of Proficiency (STAMP) created previously at CASLS. The CAP project introduces a new item development process and a new delivery algorithm for the reading section.

Native speakers of Persian identified reading passage and wrote corresponding items, with help from CASLS staff. A comprehensive review of the test items was conducted in May 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. The most promising items were selected for pilot and field testing.

Empirical information on the items was collected through a small-scale fixed-form pilot test and an adaptive field test. A total of 26 students participated in pilot testing, and 156 students participated in field testing. Speech and writing samples were collected for those test sections, but no ratings were given. Reading data from the pilot and field tests were analyzed using a Rasch methodology. The person reliability of the reading section was estimated at .88. 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 80% accurate in identifying the students' "true" proficiency level.

1 Overview and purpose of the assessment

1.1 Construct for the CAP

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

In CAP, genuine materials and realistic language-use situations provide the inspiration for 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. 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 on page 24.

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.

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

Table 1
CASLS Benchmark Levels

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

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).

Persian CAP currently measures students up through the Transitioning Level (ACTFL Intermediate-High / ILR Level 1+). Table 2 shows a detailed description of the language construct for Persian CAP.

Table 2
Language Proficiency Measured by CAP (based on Bachman & Palmer (1996))

		Beginning	Transitioning	Expanding	Refining
Grammar	Vocabulary	knowledge of limited number of common words and cognates	knowledge of some general purpose vocabulary	knowledge of most general purpose vocabulary and common cultural references	knowledge of general purpose vocabulary and some specialized vocabulary
	Syntax	little productive ability, but may be able to recognize memorized chunks	familiarity with basic syntactic structures, but not 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 are 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 Persian, 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 US. 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 suggests that educators not use Persian CAP (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, CAP can help provide empirical data on which to base decisions.

2 Description of the assessment

Persian CAP is designed to provide a general overall estimate of a language learner's proficiency in three skills in Modern Persian. 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 CAP is not a substitute for the judgment of an experienced classroom teacher. CAP can be used effectively, however, to gauge general proficiency at the start of a course for placement purposes 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 ACTLF 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 CAP can help facilitate such comparisons.

2.1 Content and structure of the CAP

The Persian CAP consists of three sections:

- Interpretive Reading
- Presentational Writing
- Presentational Speaking

The Reading section consist of multiple-choice items and is 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 for later human scoring.[1] Although the different sections of CAP 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 CAP. This is done to give language programs the maximum flexibility in using the test. Programs can choose to use all sections of CAP outright or can choose specific sections to supplement assessment practices already in place.

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

[1] CASLS does not score speaking and writing responses, but the test delivery system gives teachers the optional choice of rating students for themselves according to a simple rubric (See Section XXX).

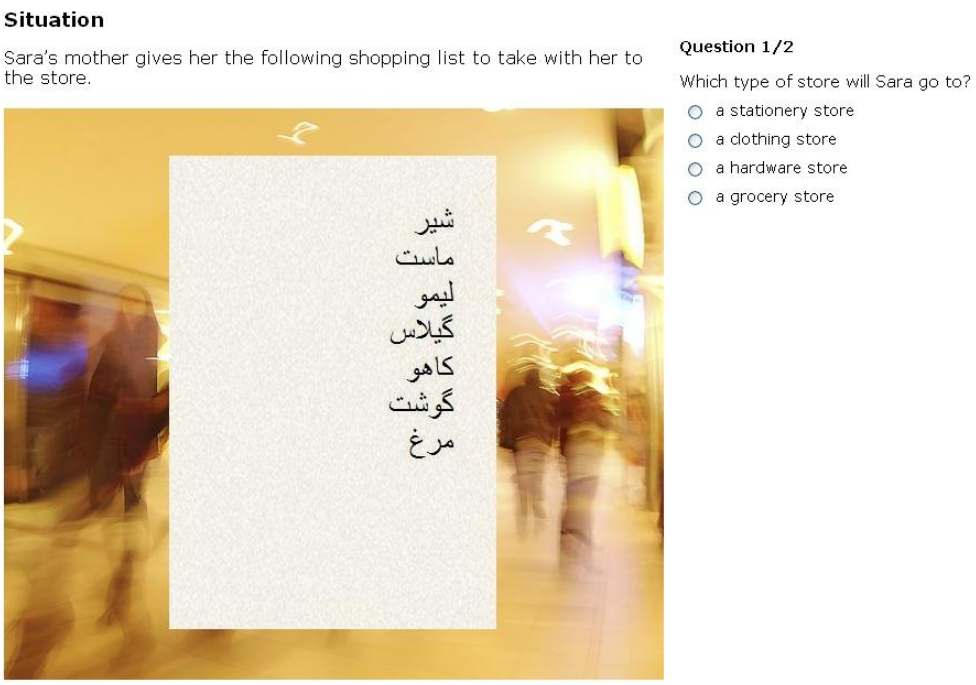

Figure 1. Persian Reading Item

2.2 Test Delivery

The Persian CAP 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 reading section of Persian CAP 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.

A visual depiction of the Persian CAP algorithm is shown in Figure 5 on page 21.

3 Test development

The general test development process for Persian CAP is illustrated in Figure 2.

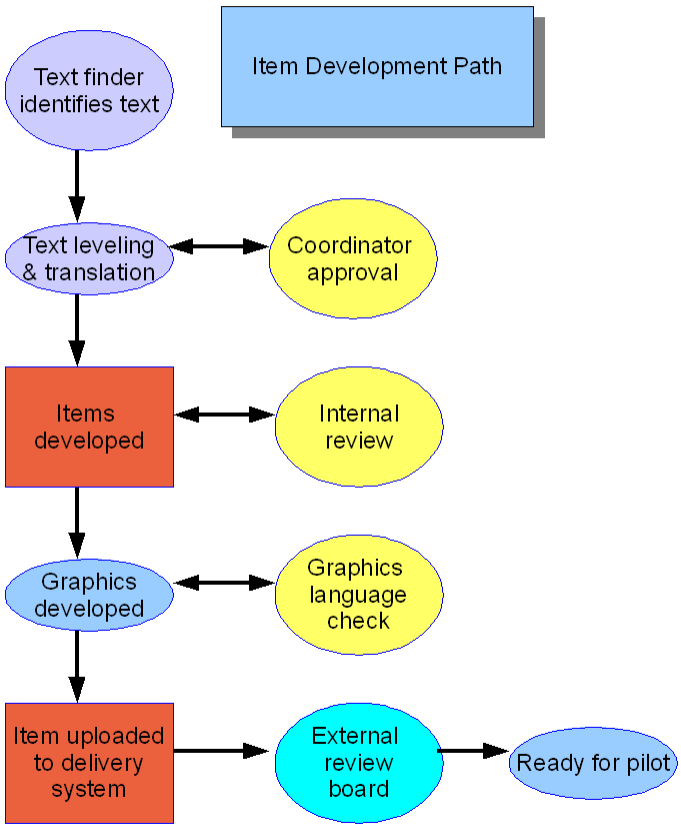

Figure 2. Item Writing Workflow

3.1 Item writing

CASLS hired three native Persian-speaking item writers to find texts and produce draft items for this project. All of the writers were Persian instructors at the university level. Two of the writers had previously participated in a passage rating workshop. All of the item writers were familiar with the ILR and/or ACTFL proficiency scales. These writers were given an overview of the item writing process and walked through the CASLS Item Writing Manual at a workshop in January 2006. Item writers also received revised manuals and more detailed item writing information and suggestions via printed material in July 2006.

The bulk of the item writing was conducted between July and December of 2006. Due to overwhelming professional commitments, one of the item writers found it impossible to both find texts and write items. Instead, the writer found and translated level-appropriate Persian texts to which items were then written by CASLS' staff.

A set of speaking and writing prompts was written by CASLS staff. As the speaking and writing prompts are delivered in English, CASLS uses similar prompts across multiple languages.

3.2 Internal review

Item writers worked remotely and submitted their items to CASLS in batches using a secure web-based interface. Items submitted by the item writers were reviewed by CASLS Assessment Director and Test Developer, and feedback was sent to the item writer. When necessary, conference calls were held to go through the reviewed items in detail. The item writers then revised and resubmitted the items. Minor, stylistic changes to the English questions and answer choices were usually made by CASLS staff without consultation. At this stage of the process, some passages and items were determined to be inappropriate for the test because of content or required background knowledge and were not developed further

3.3 Graphics development

Because the test is intended to be compatible with any computer, CASLS renders Persian text as a graphic to avoid any font display issues when the test is delivered (see sample item on page 16). For each text on the test, CASLS graphic artists imported a screenshot of the original word processor text into context appropriate images which were then uploaded to the test delivery system. A Persian-speaking graduate student was hired as an internal reviewer to ensure that the Persian text was being correctly displayed in the final item. The left-to-right nature of the Persian 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 Persian 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 Persian text in a word processing program, taking a screenshot, and remaking the graphic to incorporate the new text.

A total of 160 reading items were developed and uploaded into the CAP testing system as a result of this item development process. Four speaking and four writing prompts were also uploaded to Persian CAP.

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 compensate for the small number of Persian programs in the US.

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. A total of 164 students participated in field testing. Figure 3 shows a map of the relative number of field test participants by state.

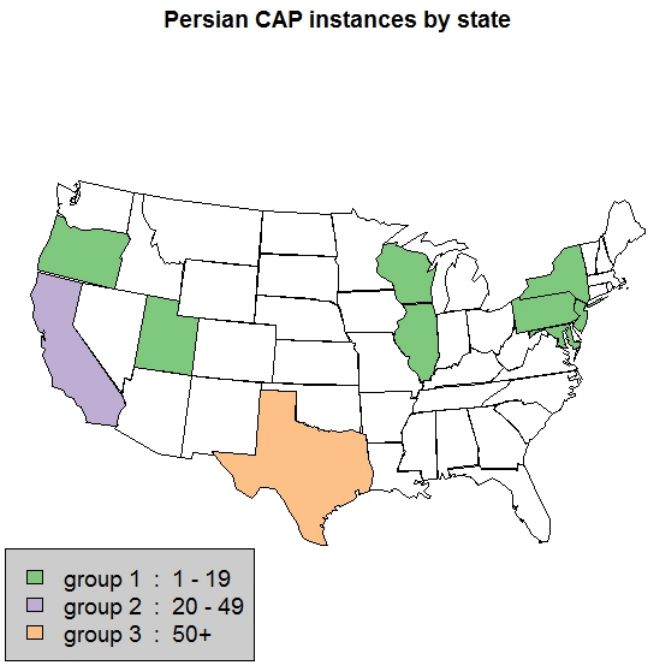

Figure 3. Map of Persian field test participants

Materials

A set of 90 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. These items were arranged into bins of 15 items across three levels of relative difficulty in a "floor first" adaptive design (See Figure 4). Since difficulty estimations were not available for these items, routing tables were created using percentage correct at level rather than item information intersections. A scoring table was created using information from the modified Angoff portion of the review session in which reviewers estimated what percent of just barely "Transitioning" level students would get a given item correct. This information was extrapolated to determine cut scores for sublevels. These scores were provided as a service to teachers to provide tentative feedback about their students.

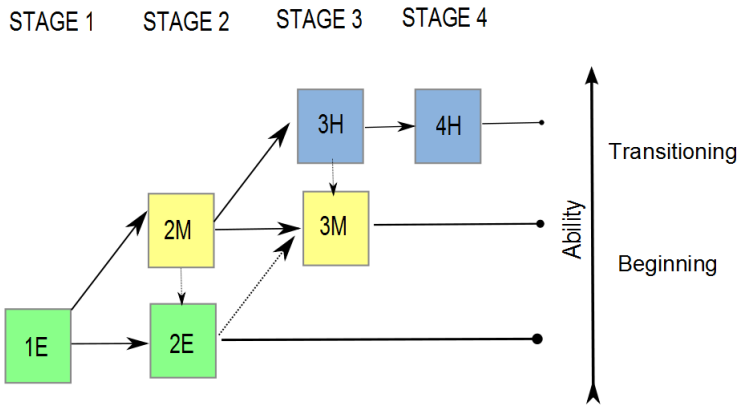

Figure 4. "Floor First" Delivery

Results

Test results were analyzed with the Rasch analysis program Winsteps (Linacre, 2008). Summary data is presented in Appendix B. The person reliability estimate was .88 and the item reliability was .93. The separation value of 2.75 indicates that the test can distinguish approximately four different levels of ability.[2] Results of the analysis were used to estimate the item difficulties for the final routing and scoring tables. One misfitting person and five misfitting responses were eliminated from the final calibration run.

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 and external reviews were used in field and pilot testing. Rasch analysis of those

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

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[3] 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 items in bins for multistage delivery. The goal was to create bins of 8 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.[4] 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 the next stage. (See Figure 5 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.

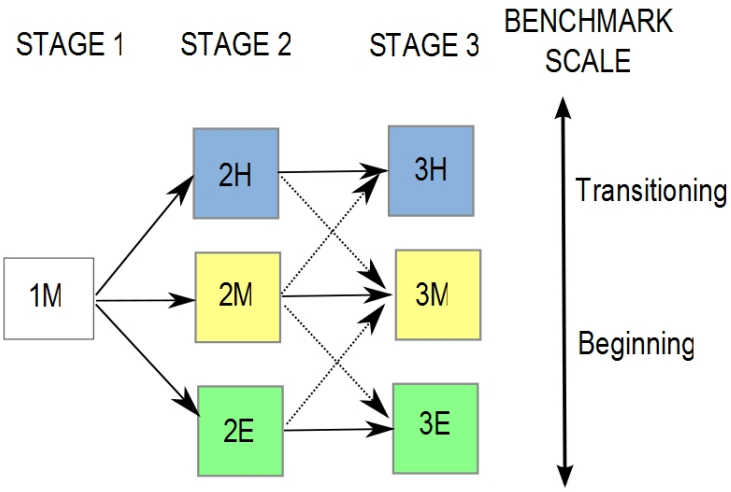

Figure 5. Delivery algorithm

4.4 Determination of cut scores

As indicated in the validity section, expert review provided an indication of the difficulty of the items in terms of proficiency levels. Cut scores were determined by calculating the median item difficulty for each major proficiency level for those items remaining in the pool. A value of 1.4

[3] A common passage with more than one associated question.

[4] For Rasch-based tests, the most informative item is one for which the test taker has a 50% probability of success.

logits was added to this value to determine the ability level needed to have an 80% probability of answering a median level question correctly. Setting the cut score to a Rasch value rather than to a particular number correct allows the particular items in the test to change while the cut score stays stable.

5 Validity evidence

5.1 External review

A comprehensive review for the reading section of Persian CAP was held at CASLS on May 2, 2008. The purpose of the review was twofold:

1) to have the quality of the items reviewed by independent experts, and

2) to provide initial evidence that the items were appropriate for the proficiency levels targeted.

Participants

The following Persian specialists participated as external reviewers in the all day session:

- Dr. Moh Ghanoonparvar (The University of Texas at Austin)
- Dr. Mahvash Shahegh
- Dr. Soheila Amirsoleimani (University of Utah)
- Dr. Mehdi Khorrami (New York University)

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

Procedure

Reviewers were first given an overview presentation of the background and design of the test and allowed to ask clarifying questions. They were then given full-color printouts of Persian CAP reading items to refer to during the review. The decision to use paper copies rather than view the items online was made primarily for primarily logistical reason. It was also felt that hard copies would be more efficient for note-taking and commenting. Once the overview of the test was completed, the reviewers were performed two different standard setting tasks.

"Big sort" For the first task, reviewers were instructed to sort items into easy, medium, and hard categories. "Easy" items were those at the lower end of the proficiency scale while "Hard" were those at the upper end of the scale, but still within the intended proficiency range of the test.[5] Reviewers tended to interpret this in terms of the familiar ACTFL and ILR scales. This sorting was done to give the reviewers an overview of all items and allow any problematic items to be discussed

[5] Items that were thought to be too difficult for the test were treated as problematic.

Table 3
Angoff Rating Correlations

	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1.00	0.34	0.38	0.44
Rater 2	0.34	1.00	0.54	0.78
Rater 3	0.38	0.54	1.00	0.71
Rater 4	0.44	0.78	0.71	1.00

prior to moving to the second phase of discussion. The big sort was done in three rounds. At first, each reviewer went through the items for that particular round individually, marking their estimated level on a master sheet. After each round, reviewers came together to discuss the items from that round.

Angoff rating After the big sort, reviewers were asked to rate the items again using a modified Angoff approach. First, the proficiency descriptions for Beginning and Transitioning levels were reviewed. Next, reviewers were asked to imagine a minimally proficient Transitioning level student. Then, for each item, reviewers were asked to estimate the probability of a minimally proficient Transitioning level student getting that item correct.[6] As with the big sort, reviewers rated the items in rounds and discussed their ratings between each round.

Results

Overall, the reviewers expressed general satisfaction with the test design and items. On a post-review questionnaire, all of the reviewers expressed confidence in the defensibility of their ratings.

The correlations between reviewers' Angoff ratings are shown in in Table 3. Rater 1 was clearly rating slightly differently from the other raters, a fact which was evident in the discussion following each of the rounds. This was partially due to the raters' expressed difficulty with conceptualizing percentages. This rater was removed from the subsequent analysis.

Ratings for the remaining three raters were analyzed using multi-faceted Rasch analysis with Facets software Linacre (2009). No standard residuals greater than 3 were observed. The "fair average" results from the Facets analysis were used to determine the cut scores for the initial adaptive version of the Persian CAP used in field test.

[6]Since probability can be difficult for some people to determine, reviewers were also given the alternative suggestion of imagining 100 minimally proficient students and estimating how many of that 100 would get the item correct.

5.2 Pilot Testing

As of this report (October 2008), Persian CAP is scheduled to begin piloting using the "Floor First" adaptive delivery system described in Section 2.2. However, a limited fixed form pilot was run in summer 2008 with students from San Diego Stage University.

Participants

Participants were students in an intensive Persian summer program at San Diego State University. No additional biodata was collected. A total of 26 subjects took the test.

Materials

From the reviewed items, a group of 74 items that had "passed" the review with no revisions was chosen to pilot. This was done since there was insufficient time between the review and the pilot to complete the revisions to the other items. These items were delivered in a non-adaptive format.

Results

An exploratory Rasch analysis was performed on the Persian pilot data. Of the 74 items on the pilot, six had extreme scores (i.e., every student got the item correct or incorrect), and these were removed from the analysis. The person reliability (analogous to Cronbach alpha) for the pilot group was estimated at .90. As can be seen from Figure 6, the Rasch estimate of item difficulties shows a trend in the expected direction, with higher proficiency items being more difficult than lower proficiency items. In addition, no examinee did better (as a percentage correct score) on Transitioning level items than Beginning level items. This did not hold strictly true for the difference between Transitioning and Transitioning-High items, but this is partially a function of the small number of items designated as Transitioning-High level on the pilot test. Based on this result, the distinction was not maintained on the field test or final version.

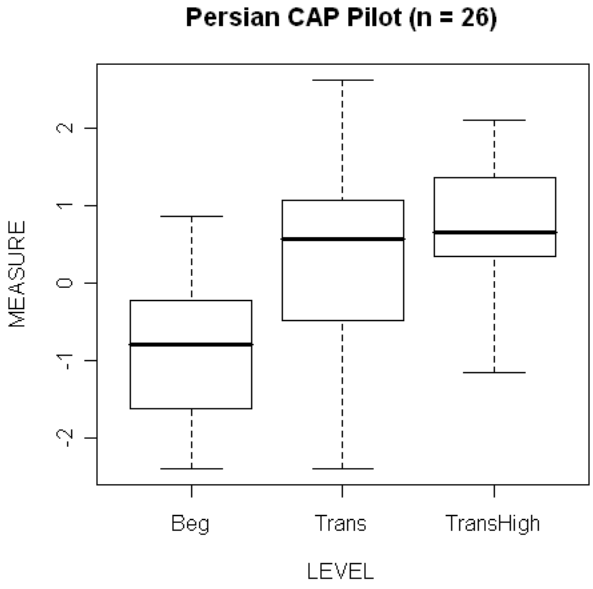

Figure 6. Persian Pilot Results

6 Score reporting

Persian CAP 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 scores

Reading scores are reported as general proficiency levels and as scaled scores. The scaled score is derived by multiplying the Rasch estimate by 45.5 and adding 500. These values were chosen to eliminate the need for decimal places in the scores. The scaled scores are simply a linear transformation of the logit scale values into a more user-friendly format and should be interpreted only in relation to cut scores for this test and not similar scores for other standardized tests. Cut scores are shown in Table 4.

Table 4
Cut Scores for Scaled Scores

Level	Reading
Beginning	370
Transitioning	609

There is approximately a ± 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

CASLS does not provide rating for the speaking or writing sections. As such, the reliability of the speaking and writing sections are unquantifiable. However, teachers are able to log in and rate their student samples based on a simple rubric. The same rubric is used for all speaking and writing items. Once rated, the average ratings across all speaking and writing items will appear on the report page. The current version of the rubric in Table 5. The relationship between proficiency levels and the possible speaking and writing scores is shown in Table 6. Teachers also have the option to view the speaking and writing responses without giving any ratings. Note that the possible scores on the writing and speaking include proficiency levels higher than those available for the reading test.

Table 5
Common Speaking Rubric

Score	Language	Score	Control
4	Speaks in multiple, clearly connected sentences. Uses a variety of sentence types and discourse organizers	4	Expansive vocabulary. Easy to understand. Tailors speech to audience. Shows awareness, though not perfect control, of discourse conventions
3	Speaks mostly in connected sentences. Uses a variety of sentence types.	3	Able to narrate in multiple time frames and express relationships (e.g., sequential, causal, etc.). Easy to understand, though may make some errors.
2	Speaks in a combination of memorized phrases and sentence-length utterances. Can occasionally string sentences together.	2	Shows evidence of original production, but may still have errors in basic structures. Generally understandable.
1	Speaks mostly in single words or memorized phrases	1	Relies on memorized elements. May be difficult to understand.
0	Little or no target language	0	Little or no target language

Table 6
Speaking Scores and Proficiency Levels

Score	Level
4.0	Refining
3.5	
3.0	Expanding
2.5	
2.0	Transitioning
1.5	
1.0	Beginning
0	

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 comptuer program.* [Version 3.68]. Chicago, IL. (http://www.winsteps.com)

Linacre, J. M. (2009). *Facets: Many-facet rasch measurement.* [Version 3.66]. Chicago, IL. (http://www.winsteps.com)

Luecht, R. M. (2003). *Exposure control using adpative 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

Persian Standard Setting Agenda -- May 2, 2008

Agenda Item	Goal	Comments	Time
Introductions	Introduce participants	Take care of any outstanding paperwork needs	8:30
CAP Overview a. Purpose b. Levels c. Sample Items d. Algorithm	Give participants an overview of the purpose of CAP, how the test will be used, what the format of items is, and how it will be delivered	Highlight proficiency versus achievement; mid-project changes vis-à-vis STAMP 2.0 project; ACTFL/ILR guidelines	9:00
Questions / clarifications	Make sure any initial concerns/questions are addressed		10:00
Standard setting intro	Give participants overview of process		10:10
Task 1 – Sorting sample items by difficulty	Raise participants awareness of text versus item difficulty; make sure participants are "on the same page"	Participants sort items individually, discuss as group	10:25
Task 2 – Initial "big" sort	Participants sort item sets into four piles (clearly Beginning, clearly Transitioning, unsure, problematic/Adv) based on passage level and overall item level (i.e., a set may have one item easier than overall level)	Participants sort individually at first, then compare results	11:00
Task 3 – Comparison	Get consensus on results		11:45
Lunch		Sweet Basil Thai restaurant	12:00
Item sorting intro	Discuss idea of "minimally proficient" intermediate student	Re-introduce ACTFL / ILR skill level descriptors	1:00
Task 4 – Item sorting (Beg/ Tran boundary)	For each item in each item set, participants indicate whether or not item would be answered by minimally proficient student		1:40
Break			2:40
Item sorting intro	Discuss idea of "minimally proficient" advanced student	Re-introduce ACTFL / ILR skill level descriptors	2:55
Task 5 – Item sorting (Tran/Expand boundary)	For each item in each item set, participants indicate whether or not item would be answered by minimally proficient student		3:20
Comparison and discussion	Sort out any outstanding issues		4:00
Evaluation	Fill out evaluation form		4:40
Wrap-up		Job well done	5:00

Figure A.1. Standard Setting Outline

B Rasch summary results

Table B.1
Persian Reading Results - Persons

Summary of 152 Measured (Non-Extreme) Persons

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	37.2	47.2	1.63	.48	.99	.1	.89	.1
S.D.	11.7	11.5	1.51	.14	.22	.9	.47	.8
Max	58.0	60.0	5.35	1.08	2.04	3.2	3.14	3.1
Min	12.0	22.0	-2.11	.32	.45	-2.2	.10	-1.6

Note. Winsteps v3.69 Table 3.1., Real RMSE=.52, TrueSD=1.42, Separation=2.75, Person Reliability=.88, Model RMSE=.50, TrueSD=1.43, Separation=2.86, Person Reliability=.89

Table B.2
Persian Reading Results - Items

Summary of 82 Measured (Non-Extreme) Items

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	65.2	82.9	.00	.41	.99	.1	.87	-.1
S.D.	34.2	37.9	1.70	.20	.12	.8	.30	.8
Max	152.0	157.0	3.87	1.03	1.33	2.8	1.50	2.5
Min	22.0	44.0	-4.04	.19	.78	-2.3	.31	-2.1

Note. Winsteps v3.69 Table 3.1., Real RMSE=.46, TrueSD=1.63, Separation=3.53, Item Reliability=.93, Model RMSE=.46, TrueSD=1.63, Separation=3.56, Item Reliability=.93

C
A
S
L
S