CASLS REPORT

Technical Report 2010-4
Unlimited Release
Printed September 2010

Supersedes NA
Dated NA

Spanish Computerized Assessment of Proficiency (Spanish 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 2010-4
Unlimited Release
Printed September 2010

Supersedes NA
Dated NA

Spanish Computerized Assessment of Proficiency (Spanish 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 Spanish Computerized Assessment of Proficiency (CAP). The development of this test was funded through the Fund for Improvement of Post-Secondary Education (FIPSE) as part of a project to investigate the use of proficiency based tests for articulation. The CAP is a proficiency-oriented test of listening, reading, writing, speaking, and contextualized grammar 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 Spanish 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 information on how the test is scored.

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

Appendix

List of Figures

List of Tables

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 Spanish Computerized Assessment of Proficiency (Spanish CAP), an online assessment of Spanish that covers a proficiency range comparable to the American Council on the Teaching of Foreign Languages (ACTFL) proficiency levels Novice through Advanced in five skills (listening, reading, writing, presentational speaking, and contextualized grammar). 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, additional skills, and a new delivery algorithm for the listening and reading sections.

Native speakers of Spanish identified listening and reading passages from authentic sources. Promising passages were translated for item development by CASLS staff and then reviewed by native speakers. Native speakers also created some reading and listening passages when appropriate authentic materials could not be located.

Empirical information on the items was collected through an adaptive field test. Over 5000 students participated in field testing. Speech and writing samples were collected for those test sections, but no ratings were given. Reading, listening and contextualized grammar data from the field test was analyzed using a Rasch methodology. The person reliability was estimated at .93 for the reading test, .89 for the listening test, and .75 for contextualized grammar. Appropriately functioning items were assembled into test panels using empirical information to establish a score table and routing table. Cut scores for proficiency levels were set at a point representing 80% probability of success for items at that level. Simulations of the delivery algorithm show a correlation of $r = .98$ between simulated test taker ability and final ability estimate on the operational reading and listening panels. The simulation also suggests that the reading and listening sections are approximately 90% 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 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. 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, listening, writing) as evidence for this capacity. An additional contextualized grammar section assesses the students' ability to distinguish between grammatically appropriate and inappropriate uses of the language.

In CAP, 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". Assessment points for the contextualized grammar section are drawn from grammatical structures typically taught in the first three years of formal language instruction.

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.

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

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

Table 1
CASLS Benchmark Levels

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

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

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

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

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

Spanish CAP currently measures students up through the Expanding Level (ACTFL Advanced / ILR Level 2). A small number of items were developed at the Refining level (ACTFL Superior), but those were not included in field testing and are not part of the operational test. Table 2 shows a detailed description of the language construct for Spanish 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 Spanish, 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 Spanish 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 additional empirical data on which to base decisions.

2 Description of the assessment

Spanish CAP is designed to provide a general overall estimate of a language learner's proficiency in four skills in Spanish, as well as Spanish grammar. 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 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 CAP can help facilitate such comparisons.

2.1 Content and structure of the CAP

The Spanish CAP consists of five sections:

- Interpretive Listening
- Interpretive Reading
- Contextualized Grammar
- Presentational Writing
- Presentational Speaking

The listening, reading, and contextualized grammar 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 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 Spanish CAP reading item may look something like Figure 1. Examinees are presented with a situation that describes a realistic language use context. A graphic contains both the Spanish text as well as contextualizing information. The test question, in English, requires the examinee to read the information in Spanish and choose the best answer from the options

[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 4).

provided. Examinees must answer the question before proceeding to the next screen. Backtracking is not allowed.

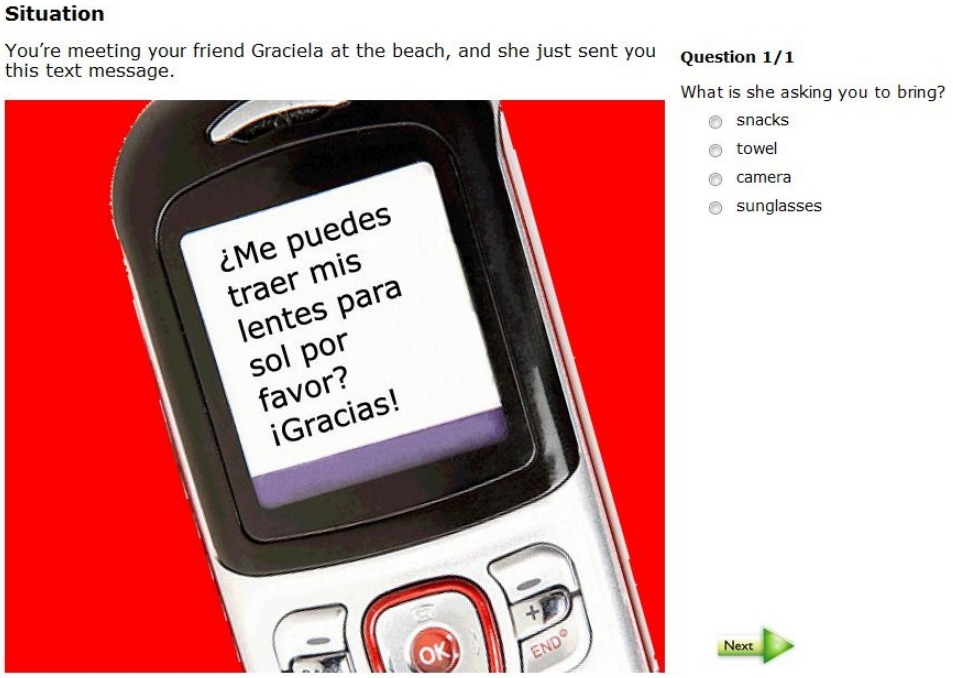

Figure 1. Spanish reading item

Spanish 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 and the playback button will no longer be active. Examinees can hear the audio only once 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 once for each of the associated items if they choose.

2.2 Test delivery

The Spanish 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 listening and reading sections of Spanish 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.

Figure 2. Spanish listening item

A visual depiction of the Spanish CAP algorithm is shown in Figure 6 on page 24.

3 Test development

The general test development process for Spanish CAP is illustrated in Figure 3.

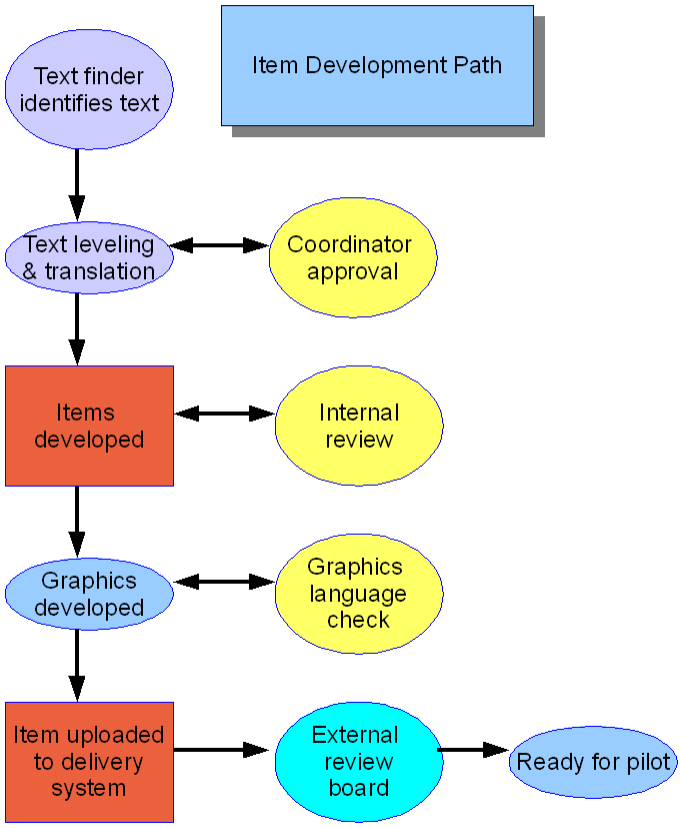

Figure 3. Item writing workflow

3.1 Item writing

CASLS hired two native Spanish-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 were 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 Spanish-speaking students came from Latin America.

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. Draft passages deemed worthy of further development were uploaded into an internal item bank database.

For advanced level texts, text finders were tasked with finding authentic listening and reading texts that best matched the test specifications and target proficiency levels. This was primarily done by searching through Spanish language resources on the World Wide Web. Many authentic texts could be discounted out of hand and 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 those texts found were kept and translated. Of those "found" texts, only a subset was considered good enough to use in item development.

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).

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.

3.2 Internal review and revisions

Throughout the item development process, items were subject to internal review. CASLS test development staff reviewed English translations of passages to ensure that the appropriate level was assigned. Staff also reviewed items and suggested revisions or additions. Finished items were reviewed by text finders to ensure that the items did indeed match the information in the passage.

3.3 Graphics development

Because the test is intended to be compatible with any computer, CASLS renders Spanish text as a graphic to avoid any font display issues when the test is delivered (see sample item on page 17). 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. The Spanish-speaking text finders reviewed finished items to ensure that the text was being correctly displayed in the final item.

A total of 220 reading and listening 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 Spanish 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 accomodate the realities of the academic schedule and give particiapant sites maximum flexibility in choosing pilot test dates.

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. Over 5000 students[2] participated in field testing. Figure 4 shows a map of the relative number of field test participants by state.

Materials

A set of 89 reading, 90 listening, and 45 contextualized grammar items were chosen for the field test. These items were chosen for having "passed" the internal reviews with no or minor revisions and for representing a broad range of topics. Items for the reading and listening sections were arranged into bins of 15 items across three levels of relative difficulty in a "floor first" adaptive design (See 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 was also constructed using simple "percentage correct at level" calculations based on the intended proficiency level of the items. These scores were provided as a service to teachers to provide tentative feedback about their students. The contextualized grammar section was delivered in a non-adaptive format.

Results

Test results were analyzed with the Rasch analysis program Winsteps (Linacre, 2008). The Winsteps commands `CUTHI` and `CUTLO` were used to cull the most off target responses for the reading test to prevent wild guesses on very difficult items by otherwise weak examinees and careless mistakes on easy items by otherwise strong examinees from unduly affecting item fit. Summary data is

[2]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. This number is a conservative estimate based on the number of tests delivered and assuming some overlap between skills.

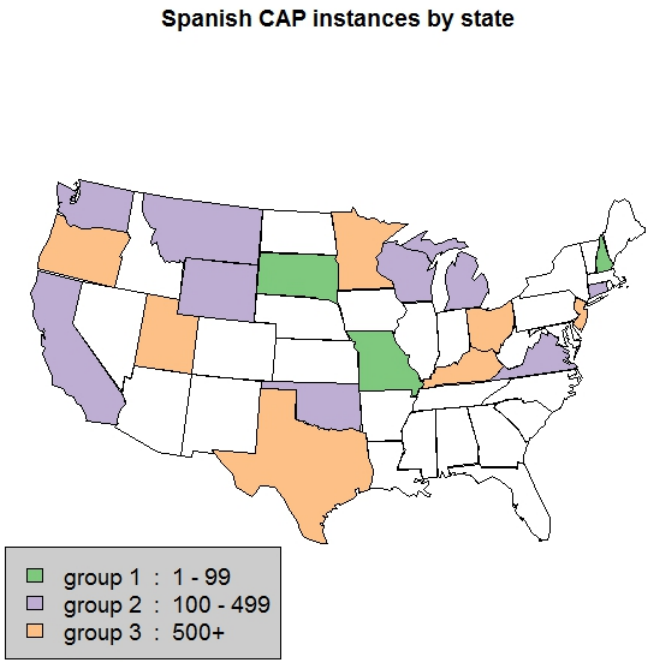

Figure 4. Map of Spanish field test participants

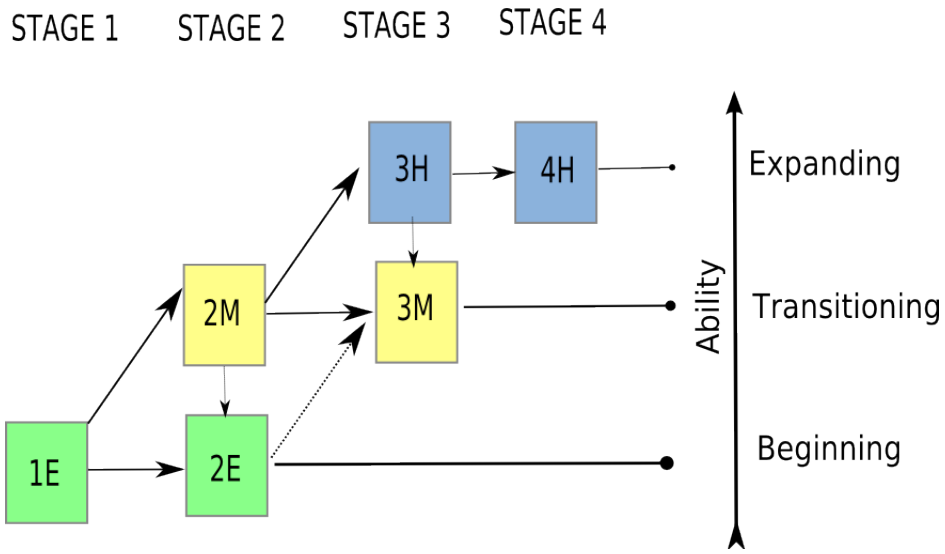

Figure 5. "Floor first" delivery

presented in Appendix A through Appendix C. In general, the items showed good fit to the model. The person separation values of 3.73 and 2.46 for reading and listening respectively indicate that the items can distinguish approximately five different levels of ability.[3] For this reason, it was determined that final score reports should focus on the major proficiency levels rather than sublevels. Results of the Rasch analyses were used to estimate the item difficulties for the final routing and scoring tables.

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[4] 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

[3] 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.

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

information about the test taker's ability given their performance on the previous bin.[5] 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 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.[6]

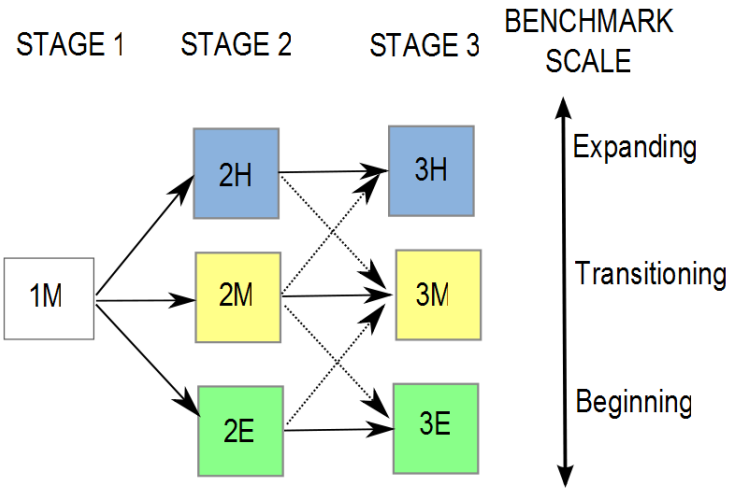

Figure 6. Delivery algorithm

For the contextualized grammar section, items remaining in the pool were arranged into a single bin of 39 items.

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. 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. The contextualized grammar section does not use any cut scores.

[5] For Rasch-based tests, the most informative item is one for which the test taker has a 50% probability of success.
[6] An example of bin information functions can be seen in Appendix D.

5 Score reporting

Spanish 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.

5.1 Reading and listening scores

Reading and listening scores are reported as general proficiency levels and as scaled scores. The scaled score is derived by multiplying the Rasch estimate by 45.5 and adding 500. These values were chosen to eliminate the need for decimal places in the scores. The scaled scores are simply a linear transformation of the logit scale values into a more user-friendly format and should be interpreted only in relation to cut scores for that particular skill on this test and not similar scores for other skills or other standardized tests. Cut scores for the various proficiency levels on this scaled score are shown in Table 3.

Table 3
Cut Scores for Scaled Scores

Level	Reading	Listening
Beginning	372	390
Transitioning	569	568
Expanding	647	637

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.

5.2 Contextualized grammar scores

Contextualized grammar scores are reported as scaled scores. As with the reading and listening sections, the score is derived by multiplying the Rasch estimate by 45.5 and adding 500. Since the conceptualized grammar items are not based on proficiency levels but rather a general sampling from the domain of grammar points typically taught in beginning classes, there are no specific cutscores for this section. There is approximately a ± 20 point standard of error for the scaled scores in this section.

5.3 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 is shown in Table 4. The relationship between proficiency levels and the possible speaking and writing scores is shown in Table 5. 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 the Refining proficiency level, which is higher than the top score possible for the reading and listening sections.

Table 4
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 5
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)

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 Rasch summary results – reading

Table A.1
Spanish Reading Results - Persons

Summary of 5690 Measured (Non-Extreme) Persons

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	28.2	41.0	.70	.42	1.00	.0	1.01	.1
S.D.	8.6	11.5	1.69	.12	.18	.9	.43	.9
Max	42.0	58.0	5.81	1.52	3.09	4.8	5.09	4.7
Min	1.0	2.0	-4.67	.32	.24	-3.2	.17	-2.1

Note. Winsteps v3.69 Table 3.1., Real RMSE=.45, TrueSD=1.63, Separation=3.59, Person Reliability=.93, Model RMSE=.44, TrueSD=1.63, Separation=3.73, Person Reliability=.93

Table A.2
Spanish Reading Results - Items

Summary of 88 Measured (Non-Extreme) Items

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	1823.8	2661.1	.00	.05	.99	-.3	.99	-.4
S.D.	1122.8	1286.2	2.00	.01	.07	3.8	.15	3.9
Max	4303.0	5652.0	4.43	.10	1.23	9.9	1.43	9.9
Min	232.0	1067.0	-4.06	.03	.88	-8.8	.74	-8.2

Note. Winsteps v3.69 Table 3.1., Real RMSE=.06, TrueSD=2.00, Separation=35.25, Item Reliability=1.00, Model RMSE=.06, TrueSD=2.00, Separation=35.55, Item Reliability=1.00

B Rasch summary results – listening

Table B.3
Spanish Listening Results - Persons

Summary of 2718 Measured (Non-Extreme) Persons

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	19.9	33.1	-.67	.44	1.00	.0	1.01	.0
S.D.	9.6	11.6	1.39	.14	.18	1.0	.34	1.0
Max	46.0	60.0	3.99	1.59	2.14	3.7	3.44	3.8
Min	1.0	2.0	5.56	.30	.36	-3.2	.31	-2.8

Note. Winsteps v3.69 Table 3.1., Real RMSE=.48, TrueSD=1.30, Separation=2.70, Person Reliability=.88, Model RMSE=.46, TrueSD=1.31, Separation=2.83, Person Reliability=.89

Table B.4
Spanish Listening Results - Items

Summary of 90 Measured (Non-Extreme) Items

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	603.4	1004.1	.00	.11	.98	-.3	.98	-.3
S.D.	557.3	841.3	1.57	.06	.09	2.7	.16	3.0
Max	2010.0	2691.0	2.52	.29	1.20	7.4	1.41	7.2
Min	42.0	80.0	-4.26	.05	.82	-5.5	.63	-5.6

Note. Winsteps v3.69 Table 3.1., Real RMSE=.13, TrueSD=1.57, Separation=12.44, Item Reliability=.99, Model RMSE=.12, TrueSD=1.57, Separation=12.54, Item Reliability=.99

C Rasch summary results – contextualized grammar

Table C.5
Spanish Contextualized Grammar Results - Persons

Summary of 2153 Measured (Non-Extreme) Persons

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	15.0	40.3	-.64	.41	.99	-.1	1.00	.0
S.D.	7.9	10.7	.93	.18	.17	.9	.27	1.0
Max	43.0	45.0	3.64	2.08	1.67	3.8	1.79	3.7
Min	1.0	2.0	-4.19	.32	.15	-3.1	.15	-2.8

Note. Winsteps v3.69 Table 3.1., Real RMSE=.46, TrueSD=.80, Separation=1.73, Person Reliability=.75, Model RMSE=.45, TrueSD=.81, Separation=1.80, Person Reliability=.76

Table C.6
Spanish Contextualized Grammar Results - Items

Summary of 45 Measured (Non-Extreme) Items

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	735.1	1951.2	.00	.05	1.01	-.5	1.02	-.2
S.D.	356.6	132.2	.88	.01	.12	3.6	.19	3.5
Max	1676.0	2129.0	1.66	.07	1.41	9.9	1.73	9.9
Min	236.0	1776.0	-2.16	.05	.81	-8.3	.76	-7.6

Note. Winsteps v3.69 Table 3.1., Real RMSE=.06, TrueSD=.88, Separation=15.45, Item Reliability=1.00, Model RMSE=.05, TrueSD=.88, Separation=15.88, Item Reliability=1.00

D Bin information

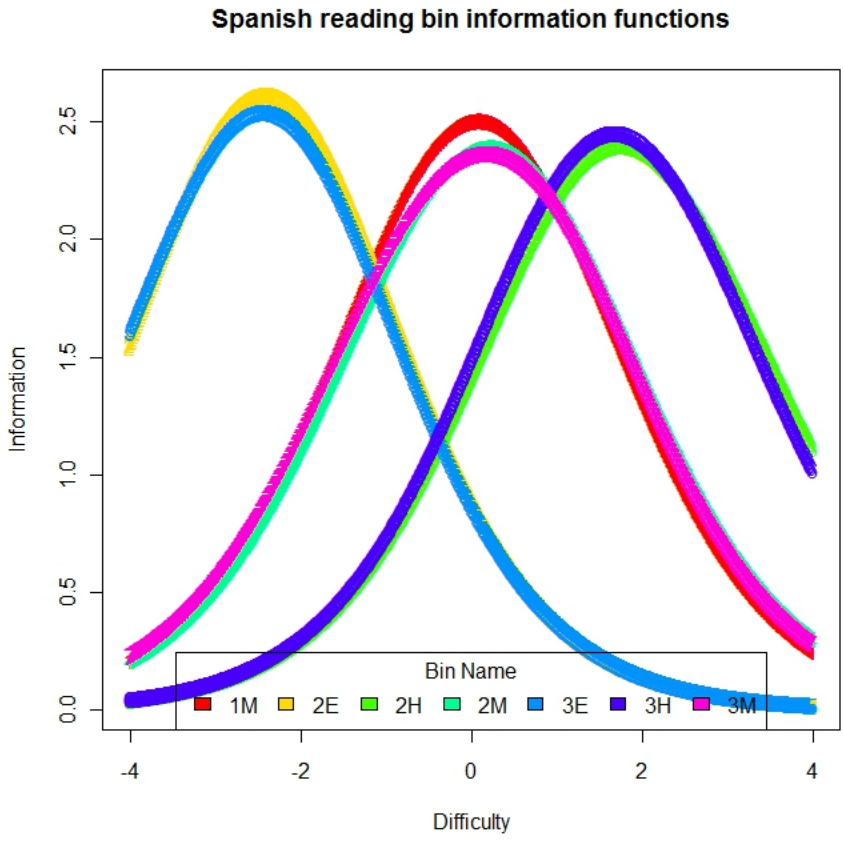

Figure D.1. Example of bin information functions used in test assembly


CASLS


FIPSE