CASLS REPORT

Technical Report 2010-8
Unlimited Release
Printed December 2010

Supersedes Hindi Final Report
Dated October 23, 2006

Hindi Computerized Assessment of Proficiency (Hindi CAP)

Sachiko Kamioka, Assistant Director
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 2010-8
Unlimited Release
Printed December 2010

Supersedes Hindi Final Report
Dated October 23, 2006

Hindi Computerized Assessment of Proficiency (Hindi CAP)

Sachiko Kamioka *
Assistant Director

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 Hindi Computerized Assessment of Proficiency (CAP). The development of this test was initially funded by the Center for South Asia Language Resource Center (SALRC) at the University of Chicago. Additional funding was obtained 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, 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 Hindi 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.

*No longer at CASLS

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

Native speakers of Hindi identified or created listening and reading passages, and created test items with help from CASLS staff. CASLS graphic artists developed images to accompany the items, and CASLS technical staff developed a test engine based on existing technology.

Empirical information on the items was collected through an adaptive field test. Over 500 students participated in field testing for the reading items. Rasch analysis of the data showed a person reliability of .90 and and item reliability of .99 for the reading section. The best items were placed into a final reading panel. Simulations on the reading panel indicate that the test is approximately 89% accurate in placing simulated students into their "actual" major proficiency level. Speech and writing samples were collected for those test sections, but no ratings were given. Items in the listening section have not been piloted.

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.

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.

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

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

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).

Hindi CAP currently includes items 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 Hindi 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 Hindi, 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 Hindi 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

Hindi CAP is designed to provide a general overall estimate of a language learner's proficiency in four skills in Hindi. 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 feedback. 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 Hindi CAP consists of five 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 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 reading item on the Hindi 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 Hindi text as well as contextualizing information. The test question, in English, requires the examinee to read the information in Hindi 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 5).

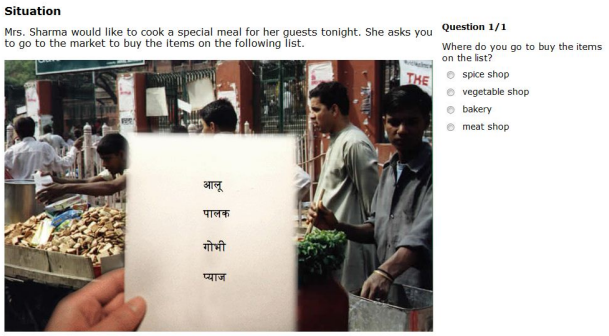

Figure 1. Hindi reading item

Figure 2. Hindi listening item

Hindi 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 Hindi 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 Hindi CAP are intended to be 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 Hindi CAP algorithm is shown in Figure 5 on page 21.

3 Test development

The content for Hindi CAP was created in two separate phases due to the nature of the funding sources. Items covering the Beginning and Transitioning levels (ACTFL Novice and Intermediate) were developed by CASLS staff and partners over a multiyear period between 2005 and 2010. Two item writers, Professor Gabriela Nik Ilieva from New York University and Professor Rakesh Ranjan from Emory University, accompanied by Steven Pulous, the director for SALRC, visited Eugene in the spring of 2005 to develop Hindi benchmarks and receive item writer training from CASLS staff. During the workshop, Hindi reading benchmarks were drafted. The benchmarks were finalized after receiving feedback from various Hindi experts in the field. A sample of the CASLS Benchmarks upon which these original items were developed is presented in Appendix A. Each item went through multiple sessions of quality checking by the project coordinator and her assistant.

In 2006, CASLS obtained additional funding to develop Expanding and Refining levels of the test. This development coincided with a reworking of the entire assessment framework including test design and delivery.[2] The development process for this most recent phase of test development is illustrated in Figure 3. Major components of this process are detailed below.

3.1 Item writing

CASLS worked with two native Hindi-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 native Hindi-speaking students came from India.

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 Hindi 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).

[2]Detailed test and task specifications are available on the CASLS website at `http://www.casls.uoregon.edu`.

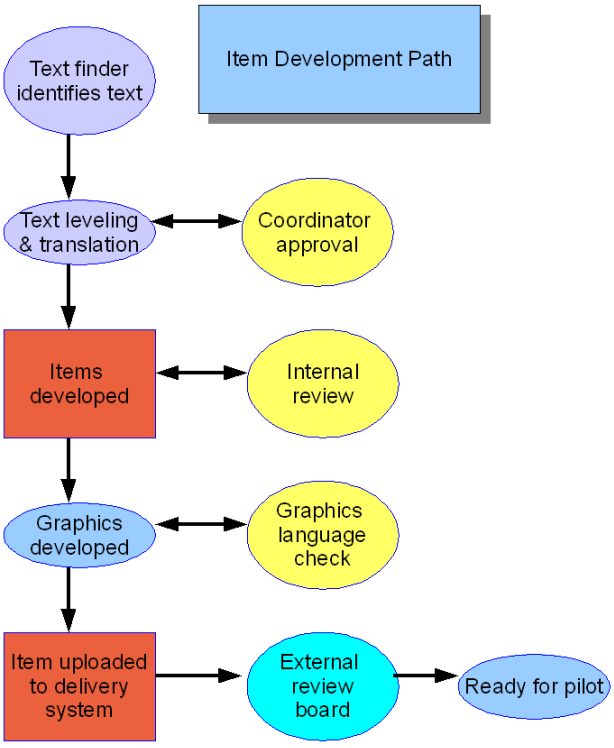

Figure 3. Item writing workflow

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

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 Hindi 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. The Hindi-speaking text finders reviewed finished items to ensure that the text was being correctly displayed in the final item. This process did not catch all errors, however, as indicated below.

3.4 External review and revisions

Throughout the process, various Hindi educators saw versions of the items in development. Several concerns were raised about the quality of the Hindi text in the reading item images. In some cases, the original Hindi text had spelling mistakes which were then reproduced in the item graphic; in other cases, Hindi characters used in word-processed documents were not always rendered correctly when those documents were opened with different versions of the software. Those mistakes were then unwittingly captured in the screenshots used for the item graphics. These problematic texts were revised and the graphics re-uploaded to the system.

A total of 454 reading and listening items were developed and uploaded into the CAP testing system over the course of several years. Four speaking and four writing prompts were also uploaded to Hindi CAP.

4 Technical characteristics

4.1 Field testing

Field testing was conducted over a multiyear period as items became available. 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. Correcting the problematic Hindi texts also took time. Results of a non-adaptive field test with lower level items is reported in Appendix B. This section describes the analysis of test data on the set of items delivered in a second field test through June 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. Just over 500 students participated in field testing of reading items. Figure 4 shows a map of the relative number of field test participants by state.

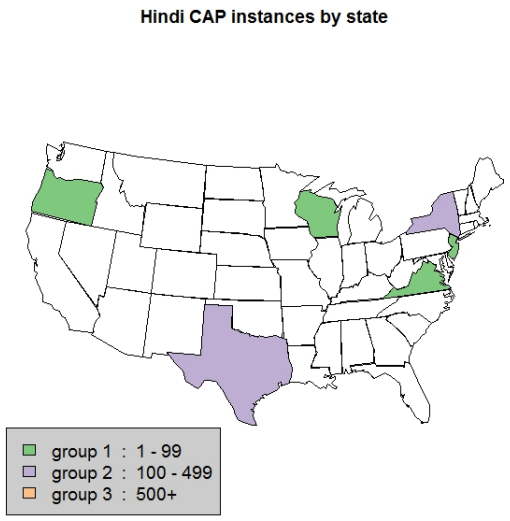

Figure 4. Map of Hindi field test participants

Materials

A set of 83 reading and 90 listening items were chosen for the adaptive 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 were arranged into bins of approximately 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.

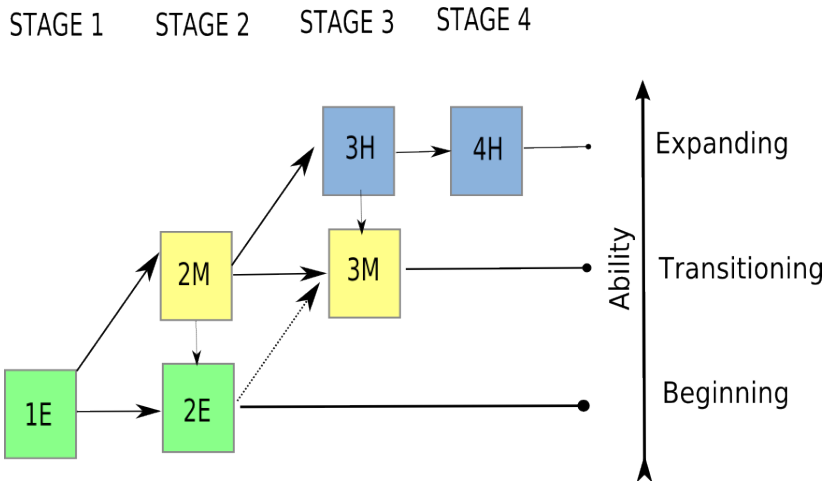

Figure 5. "Floor first" delivery

Results

Test results from the reading section were analyzed with the Rasch analysis program Winsteps (Linacre, 2008). The Winsteps commands `CUTHI = 3.0` and `CUTLOW = -3.0` were used to eliminate off-target responses. Summary data is presented in Appendix C. In general, the items showed good fit to the model. The person reliability estimate was .90 and the item reliability was .99. The separation value of 2.95 indicates that the test can distinguish approximately four levels of ability.[3] For this reason, proficiency sublevels are not reported directly. Results of the analysis were used to estimate the item difficulties for the final routing and scoring table for the reading section. Two misfitting responses were eliminated from the final calibration run.

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

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 the final field testing. Rasch analysis of those test results produced difficulty estimates for the reading items. Items with mean squared infit values between .5 and 1.5 were considered acceptable for inclusion in the item pool. In some cases, this meant that not all items in an item set were included in the operational pool. The difficulty values of these items should 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 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.[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 succesfully 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 item sets cannot be split across bins. (See Appendix D for the information functions per bin for the reading test.)

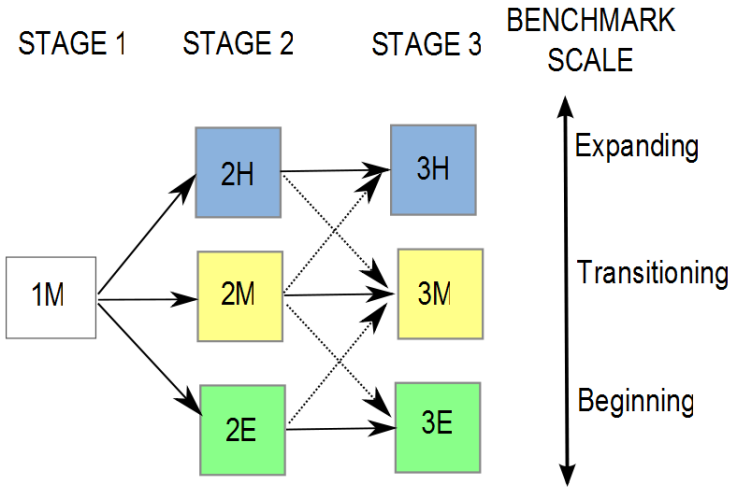

Figure 6. Delivery algorithm

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

5 Score reporting

Hindi 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 scores

As indicated in the previous section, internal and external review provided an indication of the difficulty of items in terms of desired proficiency levels. Cut scores for the reading section were determined by calculating the median item difficulty for each major proficiency level for those items 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 tes to change while the cut score remains stable. Thus, students who see more difficult items during the test will get a higher score than students who see easier items even if their numer correct score is the same. A simulation study with $10,000$ virtual students indicates that the test is about 89% accurate in identifying the students' "true" proficiency level.

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 in other standardized tests. Cut scores are shown in Table 3.

Table 3
Cut Scores for Scaled Scores

Level	Reading
Beginning	363
Transitioning	580
Expanding	643

In addition, within each major level, scores are classified as C, B, or A to give further indication of where in the level the performance was located. This should not be taken as analogous to specific proficiency sublevels. There is approximately a ± 20 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 Listening scores

Listening scores are reported as estimated proficiency levels and as benchmark scores. The estimated proficiency levels are derived from the intended level of the items. Within each major level, scores are classified as C, B, or A to give further indication of where in the level the performance was located. A test taker must be achieve the percent correct in Table 4 to be considered at that level.[5] Thus, a test taker must get at least 60% of the items correct at Transitioning level (and 90% correct at Beginning level) to be classified "Transitioning (B)" in the score report.

Table 4
Percent Correct Needed

	C	B	A
Level -1,-2	0.9	0.9	0.9
Level	0.3	0.6	0.8
Level +1			0.3

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 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 the Refining proficiency level, which is higher than the top score possible for the reading and listening sections.

[5]Note that items were written to correspond to the general levels of Beginning, Transitioning, and Expanding and not each individual sublevel.

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)

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 Sample reading benchmarks

Table A.1
Benchmark III: novice-high

Content	Context/type	Function
All of the previous plus:	Brochures	Scan for gist
Community	Maps	Extract detail
Daily routines	Simple songs	
School	Bills	
Stores/shopping	Tickets	
	Cartoons	

Table A.2
Benchmark IV: intermediate-low

Content	Context/type	Function
All of the previous plus:	Postcards	Scan for gist
Health	Letters and e-mail	Extract detail
Occupations	Invitations	
Celebrations/holidays	Announcements	
Travel/vacations	Simple narratives	
Transportation	aphorisms and proverbs	
	descriptions	

B Hindi pilot analysis

In 2007, the reading items that had been created in collaboration with SALRC were piloted at various postsecondary institutions. These items were developed according to CASLS Benchmarks. Items were piloted in a non-adaptive format, though the items were displayed in random order. The Hindi reading pilot data were analyzed[6] using Item Response Theory. Each student's proficiency measure (ability measure) was estimated relative to the other students taking the pilot. All proficiency measures were placed on a scale from 0 to 100, with an average of 50. The minimum measure was 19.3 and the maximum was 85.7. The average measure for each class is shown in Table B.3.

[6]This analysis and summary was performed by CASLS Research Director, Linda Forrest.

Table B.3
Average Proficiency Measures by Class for Hindi Reading Pilot

University	Program Year					
	IRT	(Count)	IRT	(Count)	IRT	(Count)
Columbia University	Year 1		Year 2		Year 3	
	54.7	(13)	59.0	(9)	63.5	(7)
Duke University	First Year		Second Year			
	47.5	(27)	53.1	(12)		
Emory University	Year 1		Year 2			
	44.7	(34)	56.6	(11)		
Fayetteville State University	Year 1					
	46.0	(2)				
Indiana University			Intermediate		Advanced	
			49.6	(8)	67.7	(2)
New York University	Year 1		Year 2		Year 3	
	52.8	(18)	58.2	(29)	66.1	(4)
Syracuse University	Year 1		Year 2			
	49.0	(9)	54.8	(6)		
University of California	Year 1		Year 2			
	61.1	(5)	60.4	(16)		
University of Chicago	Year 1		Year 2			
	54.0	(10)	61.6	(8)		
University of Florida	101		102		Advanced	
	49.9	(24)	53.1	(17)	56.4	(5)
University of Michigan	Year 1		Year 2		Year 3	
	51.5	(20)	58.1	(16)	71.3	(5)
University of North Carolina	Year 1		Year 2		Year 3	
	48.8	(40)	59.4	(15)	53.3	(19)
University of Texas	First Year		Second Year		Third Year	
	52.5	(48)	56.7	(6)	70.5	(11)
University of Washington	Elementary		Intermediate		Advanced	
	46.9	(35)	58.3	(11)	65.5	(4)
University of Wisconsin	Year 1		Year 2			
	56.4	(12)	55.4	(50)		
	1 semester		3 semester		5 semester	
	45.3	(17)	53.1	(8)	60.1	(8)

C Rasch summary results – reading

Table C.4
Hindi Reading Results - Persons

Summary of 498 Measured (Non-Extreme) Persons

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	22.3	34.4	.65	.44	1.00	.0	1.00	.0
S.D.	7.3	10.0	1.48	.13	.17	.9	.36	.9
Max	34.0	52.0	3.58	1.20	1.57	2.8	2.78	3.2
Min	2.0	4.0	-4.18	.33	.50	-2.5	.34	-1.9

Note. Winsteps v3.70 Table 3.1., Real RMSE=.48, TrueSD=1.40, Separation=2.95, Person Reliability=.90, Model RMSE=.46, TrueSD=1.41, Separation=3.08, Person Reliability=.90

Table C.5
Hindi Reading Results - Items

Summary of 80 Measured (Non-Extreme) Items

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	138.8	214.7	.00	.20	.98	-.2	.97	-.2
S.D.	99.2	131.4	1.86	.07	.12	1.7	.20	1.7
Max	359.0	492.0	3.67	.43	1.44	6.8	1.68	6.0
Min	15.0	59.0	-4.15	.11	.80	-3.1	.60	-3.1

Note. Winsteps v3.70 Table 3.1., Real RMSE=.21, TrueSD=1.85, Separation=8.69, Item Reliability=.99, Model RMSE=.21, TrueSD=1.85, Separation=8.79, Item Reliability=.99

D Bin information

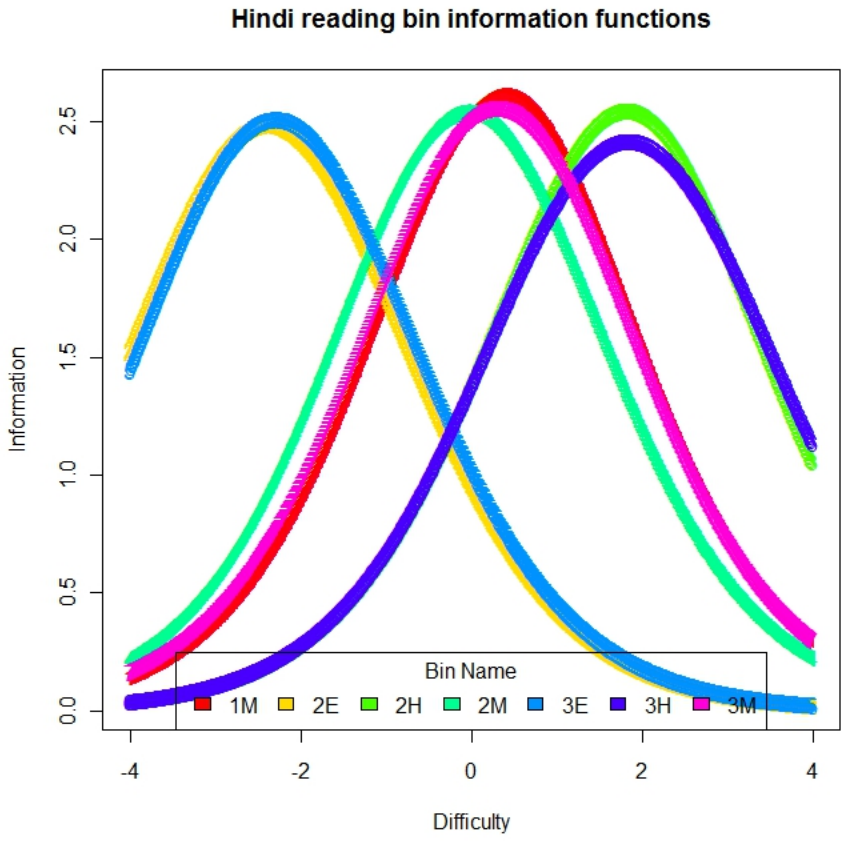

Figure D.1. Reading bin information functions used in test assembly


CASLS


FIPSE