The Online Proficiency-based Reading, Listening, and Integrated Writing External Assessment Program for Russian: A Report to the Field[1]

Saodat I. Bazarova, Maria D. Lekic, Camelot Marshall

Rising interest in the U.S. in the study and teaching of Russian language and culture, reported elsewhere in the present volume (Davidson & Garas, 2009), has coincided with the reaffirmation of the status of Russian by U.S. government agencies as a critical language. Russian is identified, for example, in the 2006 National Security Language Initiative (NSLI) as a "critical-need language," for which the production of greater numbers of advanced-level speakers is deemed essential (U.S. Department of Education, 2008, p. 1). According to the most recent ADFL/MLA report on language enrollments in two- and four-year U.S. institutions of higher education, just over one quarter of the 24,096 U.S. college and university students studying Russian were enrolled in advanced undergraduate courses in the fall of 2006 (Furman, Goldberg, & Lusin, 2007).

Investments in language over the past decades have focused primarily on training and materials development, rather than the development of assessment instruments; teachers and specialists have noted the scarcity of reliable operational testing instruments adequate for measuring student progress in mastering the language, particularly at advanced levels (Janus, 2000; Malone, Rifkin, Christian, & Johnson, 2005). To meet this growing demand for independent, proficiency-based testing at all levels, American Councils for International Education: ACTR/ACCELS (American Councils) with financial support from the Language Flagship National Security Education Program (NSEP) and the U.S. Department of Education (Title VI: International Research and Studies) has developed operational online lower-, middle-, and upper-range proficiency-based standardized testing for Russian reading, listening, and integrated writing modalities.

[1] The authors would like to extend their deep appreciation to our colleagues in the field of teaching Russian for their ongoing support with the American Councils Russian Proficiency Tests.

The new-generation proficiency-based tests in Russian replace pencil-and-paper tests produced by the Educational Testing Service (ETS) in 1987, which were also proficiency-based, standardized testing instruments. Several members of the team which contributed to the design, development, and oversight of the ETS test instruments are participating in the new Russian proficiency tests at American Councils, providing an overall comparability in approach, statistical support, scaling, and score reporting between the new tests and those of previous years. The ETS listening and reading tests existed in two forms and measured reading and listening skills within a single range: 1+ through 3 on the Interagency Language Roundtable (ILR) scale. The American Councils tests are delivered currently as online linear tests in multiple forms, three language modalities, and in three proficiency ranges: a lower-range test (0+ to 2), a middle-range test (1+ to 3) and an upper-range test (2+ to 4). As item bank and test assembly mechanisms permit, the American Councils tests will eventually be offered on demand to U.S. students and teachers as full-range computer adaptive tests (CAT) for Russian. American Councils' online test administrations are offered on specifically scheduled test dates throughout the year. Because the tests are curriculum-neutral, they may be taken by learners of Russian at any institution, level, or stage of training, heritage or non-heritage.[2] Proficiency ratings are referenced either to the ILR-scale or to American Councils' on the Teaching of Foreign Languages (ACTFL) proficiency scale.

The American Councils tests are delivered online via secured Internet connections. American Councils' multimedia team has built a delivery system which allows a secure administration of online foreign language tests anywhere in the world.[3] The program maintains a separate server designated specifically for language testing. The security of the tests is further enhanced by the presence of onsite proctors who check test-takers' identification documents and monitor testing sessions.[4]

Each American Councils Russian proficiency test is divided into three sections, which are normally taken during the same three-hour session, but may be scheduled over several days to accommodate existing school or university

[2] Native speakers are individuals whose first or primary language is Russian, and were educated principally in Russian. Heritage speakers are speakers of Russian who were born and educated in the U.S., and learned Russian at home. They differ from native speakers in that they were not formally educated in Russian. See also A. Brown (2009).

[3] While the exams can be administered online anywhere in the world, the administration is contingent upon online capabilities in the country of administration.

[4] Lesser security features can be set, depending on the specifications of the exam being administered.

class schedules. The reading and listening sections of the test are given in multiple-choice format, whereas students' writing skills are measured in the Integrated Written Communication section in a constructed response format. Student written samples are collected and graded individually by certified raters. Proficiency levels of the multiple-choice and constructed response sections range from 0+ to 4 on the ILR scale, or Novice-High through Distinguished on the ACTFL scale.

Reading and listening stimulus passages are selected from authentic materials in Russian, either from publications, TV, radio, Internet, or other media in accordance with copyright laws. Passages are selected by a team of content developers whose primary goal is to identify potential reading and listening testing materials, and to align them according to difficulty levels, text types, and themes with the ACTFL or ILR proficiency guidelines. Content developers are also trained to use special software to capture and digitize audio materials. In some cases, voices of Russian native-speaker consultants are used to produce audio files. A group of trained item writers are then instructed to generate test items for each approved reading and listening passage. These items are then reviewed both by native speaker consultants for acceptability and authenticity, and then by expert item reviewers for difficulty level, construct validity, and overall appropriateness. The test development team, consisting of Russian language and assessment specialists, reviews and approves test items for inclusion in all future American Councils Russian proficiency tests.

The present study will provide an overview of the American Councils Russian proficiency test design, as well as results of field-testing of an upper-range test. This report will help to establish the test's reliability and validity as a model test suitable for addressing a variety of assessment needs within the Russian field.

TEST CONSTRUCT

The American Councils Russian proficiency tests are based on the overall design of the Prototype AP® Russian Language and Culture Exam, which has been in use in the U.S. since 2002. The Prototype AP® Russian Exam provides a set of measurements of functional proficiency in Russian in the ACTFL lower proficiency ranges (0+ - 2). Student results on the examination are used as a predictive assessment and placement tool by American colleges and universities for entering freshmen with prior study of Russian.[5] The target population for the

[5] The design of the Prototype Advanced Placement (AP®) Exam was developed by the American Council of Teachers of Russian (ACTR) with a team consisting of Dan E. Davidson, Maria D.

Prototype AP® Russian Exam is secondary school students who have completed at least three years of high school Russian, with their final year of Russian study equivalent to that of a college-level Russian course. Heritage speakers of Russian are also encouraged to take the Prototype AP® Exam. The target population for American Councils Russian proficiency tests include high school students, college students, post-baccalaureates, and graduate students, both non-heritage and heritage students alike. Designed as proficiency tests, rather than as achievement tests, neither the Prototype AP® Russian Exam nor the American Councils proficiency tests depend on a specific curriculum or textbook. They include measures of student performance, rather than being limited to measures of knowledge. With focus on proficiency, the tests address the question, "How well do students perform in the language?"[6]

American Councils' test specifications reflect the standard proficiency guidelines established by the American Council on the Teaching of Foreign Languages (ACTFL) for the Novice-Mid to Advanced Level proficiency ranges (lower-range test); for the Intermediate-High to Superior Level (middle-range test); for the Advanced-High through Distinguished Level (upper-range test).[7] Consistent with the National Standards for Russian in the *Standards for Foreign Language Learning in the 21st Century*[8], the performance targets correspond to different proficiency levels in three modes (presentational, interpretive, and interpersonal) and in the six domains (comprehensibility, comprehension, language control, vocabulary, cultural awareness, and communication strategies). The Prototype AP® Russian Exam reflects the College Board's World Languages Framework, which represent a consensus among language professionals on curriculum, assessment, and measurement instruments, applied

Lekic, and Camelot Marshall in consultation with Bernard Spolsky (Bar-Ilan University, Israel and Gary Buck (University of Michigan); the development of the Russian Prototype Examination has also benefitted greatly over time from recommendations and contributions of Howard Everson (CUNY), Jane Rogers (University of Connecticut), Charles Stansfield (SLTI), and Hariharan Swaminathan (University of Connecticut).

[6] Maria D. Lekic. "The Russian AP® Examination Construct: Weighting and Integrating of L-2 Skills at the School-to-College Juncture," a paper presented at the 2005 National Convention of the American Association of Teachers of Slavic and East European Languages" (AATSEEL), Washington, D.C.

[7] Proficiency ranges may be represented on the ILR scale, depending on the purpose of the exam. For example, Flagship exams typically require reporting on the ILR scale, while the Prototype AP® Russian Exam reports on the ACTFL scale.

[8] See the "Russian-Specific Guidelines," *Standards for Foreign Language Learning: Preparing for the 21st Century*. National Standards in Foreign Language Education Project, 2006, Allen Press, Inc., 433-474.

to world languages education overall.[9] This framework is likewise reflected in all American Councils Russian proficiency tests. Test development, administration, and assessment procedures are undertaken in accordance with the *Standards for Educational and Psychological Testing*.[10]

TEST DESIGN

The American Councils proficiency tests and the Prototype AP® exam are composed of three online sections, aimed at the direct assessment of functional proficiency in reading, listening, and writing. These sections are delivered online, may be taken separately, or taken as a full multi-skill test requiring a total of approximately three hours to administer.[11] The Prototype AP® Russian Exam, as well as specified forms of the American Councils proficiency tests, also require an oral component (a fourth component) in the form of an oral proficiency interview (OPI), which is normally conducted over the telephone in advance of the online portions of the tests.

In order to ensure the security of the examinations and tests, two new forms of the Prototype AP® exam and multiple forms of the American Councils Russian proficiency tests are developed by American Councils each year.[12] American Councils' proficiency tests are assembled for each test administration from item banks of field-tested items. Item banks are refreshed regularly by American Councils, and poorly performing items, as well as items that may have become obsolete, are removed. Test items that appear on any live test have normally been field-tested within one of several target populations (current U.S. students overseas, students at cooperating colleges and summer programs, etc.). A development committee, made up of Russian language specialists at the high school and/or college levels, reviews the test forms and approves them. Stimulus passages (texts) and tasks (items) address a range of proficiency levels, as indicated in the test specifications. Stimulus materials (texts) are provided exclusively in the target language. Instructions, reading and listening multiple-

[9] *The College Board's World Languages Framework,* 2006, www.collegeboard.com.

[10] *Standards for Educational and Psychological Testing*, 1999. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. American Educational Research Association, Washington, D.C.

[11] See Lekic, M. D., Cynthia Martin, Camelot Marshall. Kenneth Petersen, "Technology Files: Secrets and Strategies. The Online AP® Russian Course and Examination," presented at the 2006 National Convention of the American Teachers of Foreign Languages (ACTFL), Nashville, TN.

[12] The exam forms are equated across exam administrations from year to year, by including in each exam form a mini-exam made up of multiple-choice items that represent the content and test and item characteristics of the earlier form.

choice questions (four-answer options) are in English.[13] Results for each section of the tests are calculated independently and reported both as raw scores and as proficiency ratings.[14] In the case of the Prototype AP® Russian Exam, each section is weighted equally in calculating the overall AP® grade, reported on an established score range of one to five.

The reading comprehension section consists of 10-18 authentic texts (short texts at the lower levels, longer texts at the advanced level and above) followed by one to three multiple-choice questions. Text types are varied and include formatted layouts (announcements, schedules, brochures), as well as standard paragraph-length texts. Topics range from personal information, simple interpersonal communications, social conventions, and routine tasks, to formal announcements, advertisements, or fliers, to more advanced topics with lengthier prose texts, such as news reports, policy statements, literary texts, opinion pieces, etc. Topics are age-appropriate for those taking the test. They address both the interpersonal and interpretive modes of the *National Foreign Language Standards* and are checked carefully for consistency with the *Fairness and Equity* guidelines published by the Educational Testing Services.

The listening comprehension test closely parallels the reading comprehension section. Authentic spoken stimulus passages (10 - 18 in number) may include oral texts, conversations, announcements, radio clips or news reports. In listening comprehension, the test-taker functions in several different roles, depending on the audio stimulus. For example, the test-taker may overhear a conversation or monologue, or may function as the addressee, if the audio is addressed to the test-taker.[15] Topics range from personal information, simple interpersonal communications, social conventions or routine tasks to formal announcements and reports over the media, to advanced topics with extended discourse, such as interviews, short lectures, speeches, and opinion pieces from the news media. As with the reading section, topics are reviewed for age

[13] Test questions are posed in English, rather than the target language, so as to focus on measurement of the candidate's comprehension of the texts and audios, rather than on any possible misunderstandings of the test questions themselves, if posed in the target language.

[14] Russian language experts attend standard setting meetings in order to determine cut scores, by which the raw scores on the multiple-choice sections can be reported in terms of proficiency levels. See later discussion on standard setting procedures.

[15] Listening tasks are often divided into three broad categories: participatory (where the hearer is actively participating (speaking) in a conversation), non-participatory (such as listening to a college lecture or a phone message), and overheard (where it is not known that someone else is listening).

appropriateness and address the interpersonal and interpretive modes of the *National Foreign Language Standards*.

The Integrated Written Communication (IWC) section of the exam is required for Prototype AP® Russian Exam test-takers as a means of assessing the students' overall language abilities. The goal of this section of the test is not to measure the student's ability to recall facts in isolation, but rather to integrate knowledge and abilities in order to apply their language proficiency in a wider context. The integrated writing section of the exam includes dialogue completion and integrated tasks.

A. In the dialogue completion portion, students are shown dialogue segments containing missing lines. The task for the test-taker is to write out the missing questions or responses which complete the dialogue in a linguistically and culturally appropriate manner.

B. The integrated tasks portion of the exam is "intermodal," and requires students to function in more than one skill in order to work through a real-life situation. The test-takers are given two scenarios, each having two prompts (listening audios and reading texts), with each prompt requiring an extended written response. For example, a scenario might begin with an e-mail text (reading), requiring a written response, followed by a voicemail message (listening), requiring another written response, with the combination creating a contextualized situation. The information that test-takers use to compose their response(s) is provided in the listening and reading texts, and students have a choice of keyboard layout (phonetic or standard) for their Russian fonts. Students are instructed in English as to what the written task entails. Both the interpersonal and presentational modes of the *National Foreign Language Standards* are reflected in this section of the test.

As noted above, rating of the written responses is performed by certified testers of Russian who are trained specifically in the application of the scoring rubrics for this examination. American Councils selects approximately 10 percent of rated responses at random for double-rating, as a quality assurance measure and a means of monitoring inter-rater reliability. In cases where ratings diverge, the response is submitted to a third-rater for an independent evaluation and final scoring.

Building on the Integrated Written Communication section of the Prototype AP® Russian Exam, the American Councils proficiency tests of Russian also include a writing section to assess proficiency in writing. Depending on the test specifications, the integrated task scenarios and dialogue completions may be more complex in order to account for the demands of higher proficiency levels.

FIELD-TESTING AMERICAN COUNCILS' RUSSIAN PROFICIENCY TEST

The American Councils Russian proficiency test, which currently serves the Russian Flagship Programs for pre- and post-program testing purposes, was field-tested among multiple cohorts of U.S. college students who were at the advanced level of Russian study, typically the third-year or higher. The field-test data used in this report was collected over the course of two years of test administration, which included participants in summer and academic year language programs, as well as domestic Russian Flagship programs.[16] Like the Prototype AP® Russian Exam, the American Councils Russian proficiency test consisted of three sections: reading comprehension, listening comprehension, and integrated written communication.[17] Since this test included middle- and upper-range proficiency measurements, the test specifications for both reading and listening comprehension are as follows:

Level 1+	3-5 items
Level 2	7-9 items
Level 2+	8-11 items
Level 3	10-13 items
Level 3+	10-13 items
Level 4	7-10 items

Two parallel forms of the test were developed and administered. Following the preliminary analyses of the field-test data and item analyses, a standard setting was held for the reading and listening comprehension sections of the test in

[16] The Russian Domestic Flagship Programs participating in the field-testing of the American Councils proficiency test between 2006 and 2009 were Bryn Mawr College, University of Maryland, Middlebury College, and UCLA. The authors wish to express their gratitude to these programs and their students for participation in the field-testing of the American Councils Russian proficiency test, and, in particular to Sharon Bain, Cindy Martin, Karen Evans-Romaine, and Olga Kagan.

[17] Although writing data were also collected from the field-testing, only the results from the reading and listening sections are presented in this report.

order to determine cut scores, which would allow raw scores to be distributed along an ordinal proficiency scale.

STANDARD SETTING FOR READING AND LISTENING MULTIPLE-CHOICE SECTIONS: APPLICATION OF THE BOOKMARK METHOD

With each newly developed proficiency test, American Councils assembles a panel of external experts for the specific purpose of determining the cut scores that mark the boundaries between proficiency levels for each form of the online test.[18] Experts from the field of Russian were selected based on their experience in teaching Russian at secondary and tertiary education settings, as well as their familiarity with the ILR proficiency scale, and were invited to participate in a three-day standard setting session. The primary goal of the standard setting session, held most recently in July 2009 at American Councils, was to determine the cutoff points for the reading and listening sections of the two forms of American Councils' Russian proficiency tests (FDHM and FDHA).[19] Standard setting participants were asked to take the reading and listening comprehension sections of FDHM and FDHA prior to the meeting, in order to become familiar with each section of the test and its online administration.

At the meeting, the panelists received a set of standard setting materials: agenda, the ILR proficiency scale descriptors, the general ACTFL proficiency guidelines, Russian-specific guidelines, hard copies of the actual online reading and listening comprehension test sections, as well as corresponding answer keys, ordered item booklets (OIB) (test items arranged by descending *p*-values) for each section, evaluation forms, a set of colored bookmarks and clips, and printed forms to record cut scores. A separate electronic file was created to track the panelists' recommended cut scores. The impact of these scores on the distribution of students' language skills on the ILR proficiency levels was presented to the group upon completion.

[18] Standard settings are held when there are a sufficient number of test-takers to provide ample data for the analyses.

[19] FDHM and FDHA are considered parallel forms of the same test. However, the test development committee decided to conduct separate standard settings for each test, due to anticipated low number of students taking FDHA. Test equating procedures are used to produce future parallel exam forms of the test, to which the determined cut scores will be applied.

THE BOOKMARK METHOD

The Bookmark procedure, a consensus-building strategy, served as the foundation for the standard-setting process. It is one of the most widely used standard setting procedures in educational assessment. This method is based on item response theory (IRT) models, which produce estimates of both item and test-taker characteristics. Depending on the test format, various IRT models are used in reordering test items for the standard-setting committee. For the dichotomously scored reading and listening sections of the tests, the Rasch model was run in the WINSTEPS statistical environment (Linacre, 2009).

As the name of the procedure implies, panelists are asked to review test items and place their markers on items that differentiate between proficiency levels. These markers later become the cut scores that determine students' performance on the tests. For the Bookmark method, all item analyses are conducted prior to the standard-setting session, and all items are aligned along the difficulty continuum and assembled into an ordered item booklet (OIB). OIBs for each test section were compiled using the criteria described by Cizek and Bunch (2007).

Field-testing of the American Councils proficiency tests was conducted at participating institutions with onsite proctors. Individual student responses were collected and merged into one data file for Rasch analyses. To facilitate panelists' tasks,[20] the resultant item difficulty measures were transformed (converted achievement level) using the following formula:

((Rasch Measure x 100) + 500)

All test items were ordered in the OIB, according to their converted values of item difficulty from lowest (the easiest item) to highest (the most difficult item), rather than the proficiency levels of the items. For each item, its corresponding text, the item stem with response choices, its converted achievement level, *p*-value,[21] and ILR item proficiency level were provided. There were a total of 38 items in the reading comprehension and 35 items in the listening comprehension sections of the tests.

[20] The Rasch model generates statistics in negative and positive values. In order to make the values easier for the panelists to work with, the values are converted to a positive value scale.
[21] The "*p*-values" is the percent of test-takers responding to an item correctly.

TRAINING OF THE EXPERT PANELISTS

The panel of experts was introduced to the Bookmark method and instructed to focus on the definition of "a borderline student," a critical notion in establishing cutoff scores in the test sections. In the context of the Russian language tests, the term "borderline student" signifies a non-native English-speaking learner of Russian who is on the cusp of shifting from one proficiency level to the next. A discussion of language characteristics at each proficiency level and their relation to the notion of a borderline student further enhanced the panelists' understanding of the task. For each cut score, the panelists applied their understanding of a borderline student at the corresponding ILR proficiency levels. The panelists were directed to place their bookmarks on the last page for which a borderline student had a 50/50 chance (RP50) of answering the item correctly. Although the target item response probability used in this standard setting session was .67 or 2/3 chance (RP67) of responding to an item correctly, it is generally easier to place cut scores when RP50 was utilized. The final recommended cut scores at RP50 generated by the panelists were averaged and adjusted to reflect a cut score at RP67.[22] The panelists were to place five cut scores: 1+/2, 2/2+, 2+/3, 3/3+, and 3+/4 on two test forms, each with a reading and listening comprehension section.

READING COMPREHENSION

The cut scores for the reading comprehension section of both tests were set in two rounds. At the beginning of the first round, the panelists reviewed the ILR guidelines on reading language skills. The first practice cut score (L1+/L2) was established under the guidance of a session moderator. The purpose of the first practice cut score was to assist the panelists in implementing the Bookmark procedure. All questions and comments about the procedure were fully addressed during the group discussion of the practice cut score. Panelists continued the process for round one until all cut scores were placed and recorded on a separate form. Results were shared with the group, and a discussion followed regarding the differences and similarities of cut scores.

[22] Rasch difficulty measures are calculated based on the 50 percent likelihood of getting an item correct. To raise the final cut score to RP67, a widely accepted target response probability used for standard settings, a value of 69 was added to the final average score of the panelists' proposed cut scores.

Outcome tables shown to the panelists included individual cut scores, the mean, median, and mode of the scores along with the standard deviation.[23]

Following the first round discussion, panelists continued with the second round of placing their cut scores, having the opportunity of revising their initial placements, or keeping them the same. The cut scores were again presented to the group. The outcome of the second round revealed that the discrepancy in cut scores, previously found in the first round among panelists, decreased significantly for some of the cut scores (see Table 1 for FDHM and Table 7 for FDHA). The average scores were then adjusted to reflect response probability value of .67 (see Tables 2 and 8). Then, using the panelists' adjusted cut scores, student performance on the reading sections were established and shared with the group for a final review (see Figure 1 for FDHM and Figure 4 for FDHA). For both reading comprehension sections, the panelists decided not to hold a third round of bookmarking, as they were confident that the results reflected those of their students they typically teach. The second round became the final round.

LISTENING COMPREHENSION

The cut scores for the listening comprehension section of FDHM and FDHA were set in the same fashion as was done for the reading comprehension. Panelists reviewed the standard-setting process and the ILR guidelines on listening language skills. They then held a guided practice round for the first cut score (L1+/L2). All questions and comments about the procedure were fully addressed during the group discussion of the practice cut score.

The panelists went on to mark the pages of the remaining cut scores. They recorded their proposed cut scores separately, and the results were shared with the panelists. The outcome tables shown to the group included individual cut scores, the mean, median, and mode of the scores along with the standard deviation. Following the discussion of the first round results, the panelists continued on to the second round of placing their cut scores, which allowed them the opportunity to change the placement of their cut scores. The results were presented to the panelists. The divergence in the individual cut scores diminished. The average cut scores were again corrected for RP67 (67 percent chance of getting an item correct), and the distribution of students' performance

[23] While complete consensus among the panel of experts is unlikely, the discussion is a way for panelists to express their reasoning behind their cut scores, reflect upon the comments from the group, and review and/or revise their initial cut score placements. From a statistical point of view, the discussion aims to reduce the variance among the panelists in their cut scores. That is, to have subsequent rounds result in smaller standard deviations.

on the test form, given in proficiency levels based on these cut scores, was submitted for the panelists' scrutiny. The panelists were satisfied with the results of the second round for the FDHM listening comprehension section (see Tables 3 and 4 and Figure 2), so round two for this particular section became the final round. However, they decided that there was a need for another round of the Bookmarking process, after viewing and discussing the round two outcomes of the FDHA listening comprehension section. The third round of placing cut scores proceeded in accordance with all previous rounds. The results of the third round were shared with the group, and then the cut scores were adjusted to RP67. The student distribution along the proficiency level continuum was then revealed to the panelists (see Tables 5 and 6 and Figure 3). During the discussion of these results, the panelists observed that the divergence in the cut scores among panelists decreased for all but the last cut score. Nevertheless, they approved these cut scores as the final ones for the FDHA listening comprehension section. Round three became the final round.

SUMMARY COMMENTS

The present study has sought to provide an overview of the American Councils Russian proficiency test design, development, and, more specifically, the results of field-testing and standard setting of the middle- and upper-range proficiency test, which document the test's construct validity as a proficiency test, suitable for addressing a variety of assessment needs within the Russian field.

As with the Prototype AP® Russian Exam construct, development, administration, analysis, and scoring, American Councils adheres to a strict protocol for all the components of the lower-, middle-, and upper-level range American Councils Russian proficiency tests that it produces. Although the relatively small numbers of annual test-takers in the less commonly taught languages makes certain statistical procedures more difficulty to perform, American Councils' testing reflects a system of multiple checks and balances to ensure quality test results in keeping with current best practices in standardized test development, administration, and scoring.

Table 1. *Proposed Cut Scores FDHM Russian Reading Comprehension: Round Two (Final)*

Panelist	Converted Item Measure for 1+/2	Converted Item Measure for 2/2+	Converted Item Measure for 2+/3	Converted Item Measure for 3/3+	Converted Item Measure for 3+/4
1	287	451	492	562	603
2	197	451	492	531	603
3	330	476	492	567	608
4	287	480	562	608	718
5	348	451	504	563	568
6	330	451	504	563	608
7	315	451	492	562	608
8	348	451	480	562	608
9	315	451	492	562	603
Mean	306	457	501	564	614
St. Dev.	46.70	11.95	23.94	19.53	40.99
Median	315	451	492	562	608
Mode	287	451	492	562	608

Table 2. *Final Cut Scores and Impact Data for FDHM Russian Reading Comprehension Section at RP67*

ILR Proficiency Level	Cut Score	Percent of Students	Cumulative Percent of Students
1+	375	0.00	0.00
2	526	20.00	20.00
2+	570	65.83	85.83
3	633	9.17	95.00
3+	683	5.00	100.00
4	---	0.00	100.00

Figure 1.
Distribution of Students in Percentages by ILR Proficiency Levels and by Heritage Information at RP67: FDHM Russian Reading Comprehension

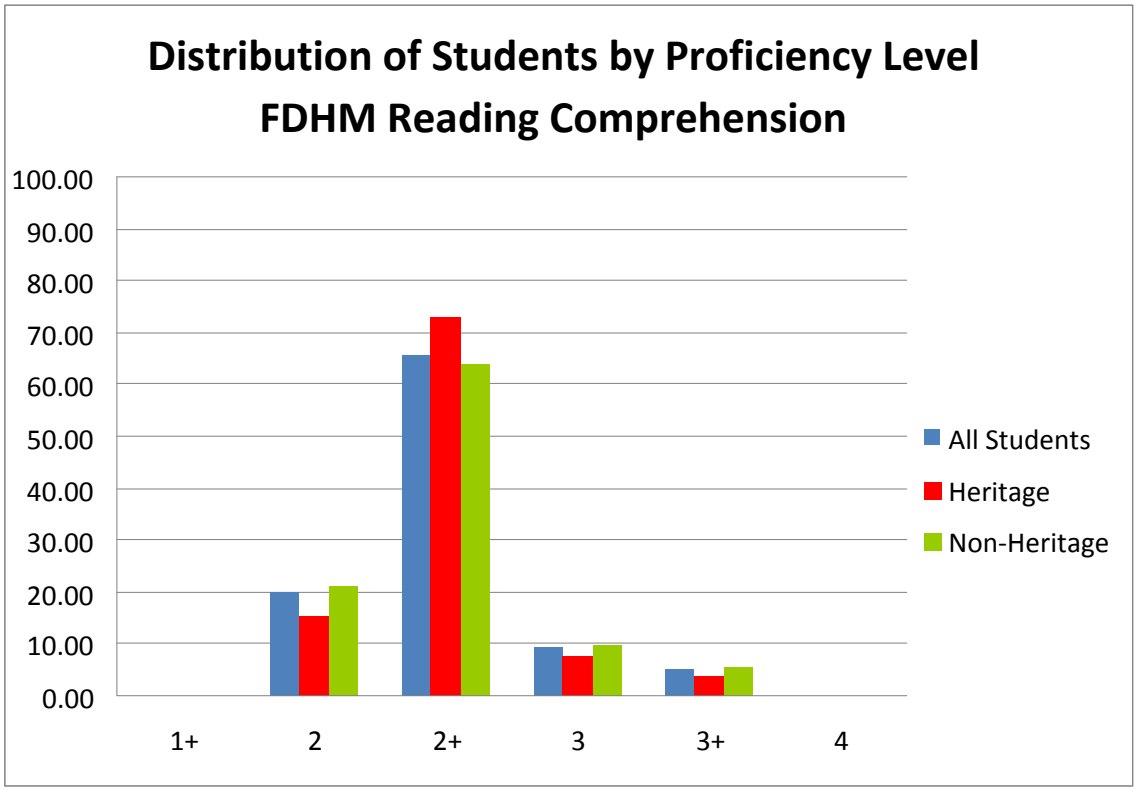


Table 3. *Proposed Cut Scores FDHM Russian Listening Comprehension: Round Two (Final)*

Panelist	Converted Item Measure for 1+/2	Converted Item Measure for 2/2+	Converted Item Measure for 2+/3	Converted Item Measure for 3/3+	Converted Item Measure for 3+/4
1	343	396	501	525	592
2	335	368	455	559	592
3	343	411	488	559	606
4	293	370	439	503	600
5	343	396	439	501	577
6	343	439	448	559	599
7	343	425	488	525	599
8	343	387	488	525	592
9	343	387	455	501	564

Mean	337	398	467	529	591
St. Dev.	16.55	23.75	24.23	24.98	13.05
Median	343	396	455	525	592
Mode	343	396	488	525	592

Table 4. *Final Cut Scores and Impact Data for FDHM Russian Listening Comprehension Section at RP67*

ILR Proficiency Level	Cut Score	Percent of Students	Cumulative Percent of Students
1+	406	0.00	0.00
2	467	20.00	20.00
2+	536	24.17	44.17
3	598	35.00	79.17
3+	660	20.83	100.00
4	---	0.00	100.00

Figure 2.
Distribution of Students in Percentages by ILR Proficiency Levels and by Heritage Information at RP67: FDHM Russian Listening Comprehension

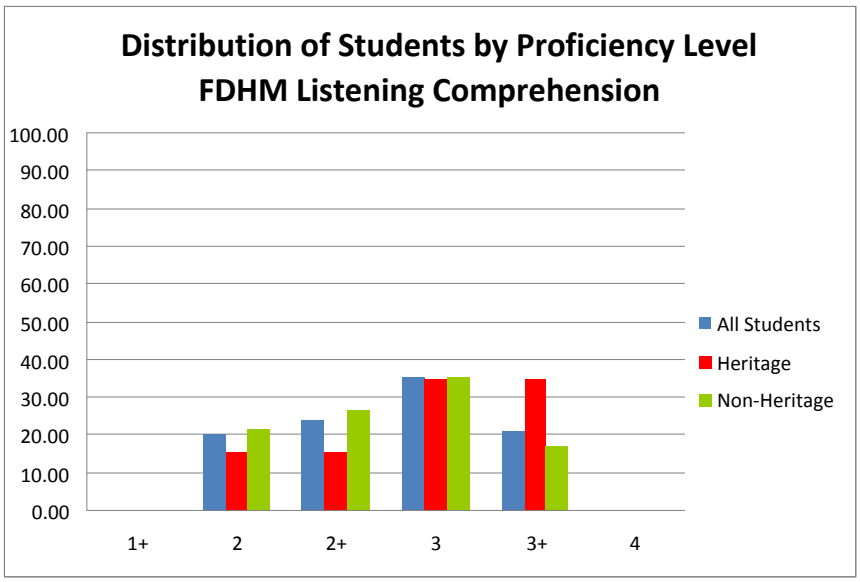


Table 5. *Proposed Cut Scores FDHA Russian Listening Comprehension: Round Three (Final)*

Panelist	Converted Item Measure for 1+/2	Converted Item Measure for 2/2+	Converted Item Measure for 2+/3	Converted Item Measure for 3/3+	Converted Item Measure for 3+/4
1	270	395	504	560	615
2	395	462	504	576	650
3	395	494	560	650	700
4	270	494	576	650	672
5	442	494	560	622	667
6	365	442	504	629	667
7	326	462	524	622	682
8	395	471	527	650	741
9	326	462	524	615	650
Mean	354	464	531	619	672
St. Dev.	59.78	31.63	27.40	32.30	35.20
Median	365	462	524	622	667
Mode	395	462	504	650	650

Table 6. *Final Cut Scores and Impact Data for FDHA Russian Listening Comprehension Section at RP67*

ILR Proficiency Level	Cut Score	Percent of Students	Cumulative Percent of Students
1+	423	0.00	0.00
2	533	28.36	28.36
2+	600	19.40	47.76
3	688	52.24	100.00
3+	741	0.00	100.00
4	---	0.00	100.00

Figure 3.
Distribution of Students in Percentages by ILR Proficiency Levels and by Heritage

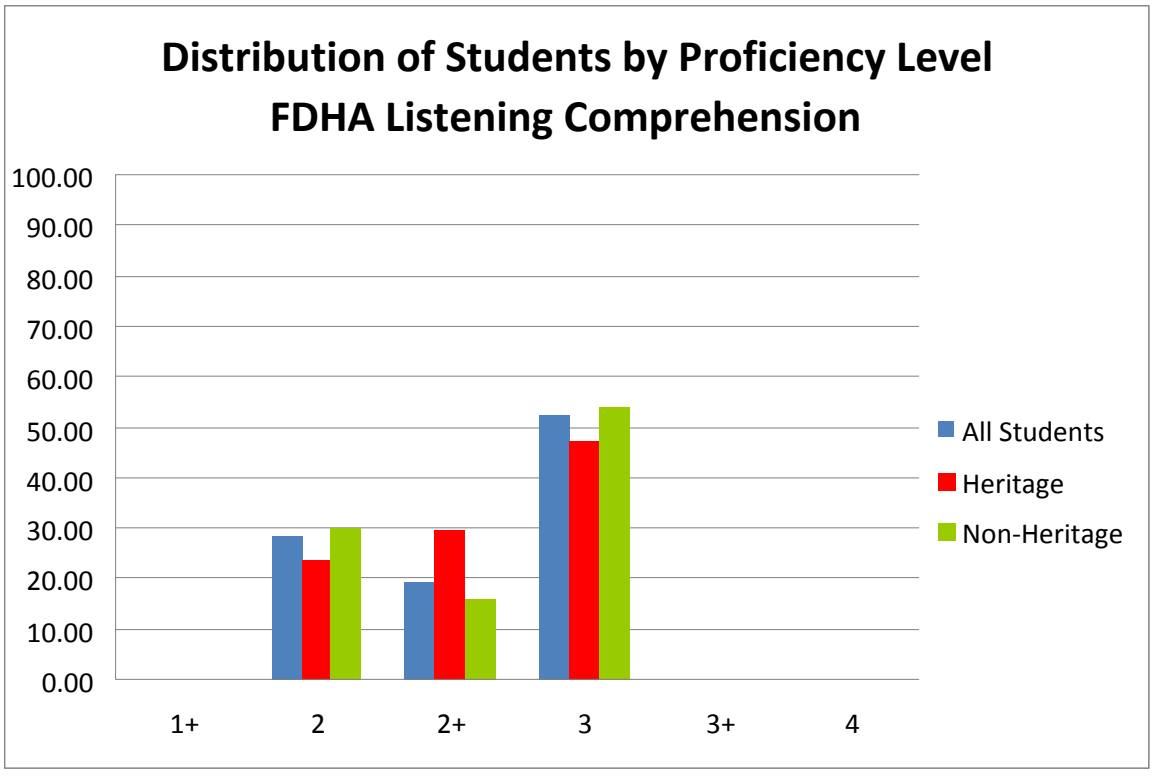


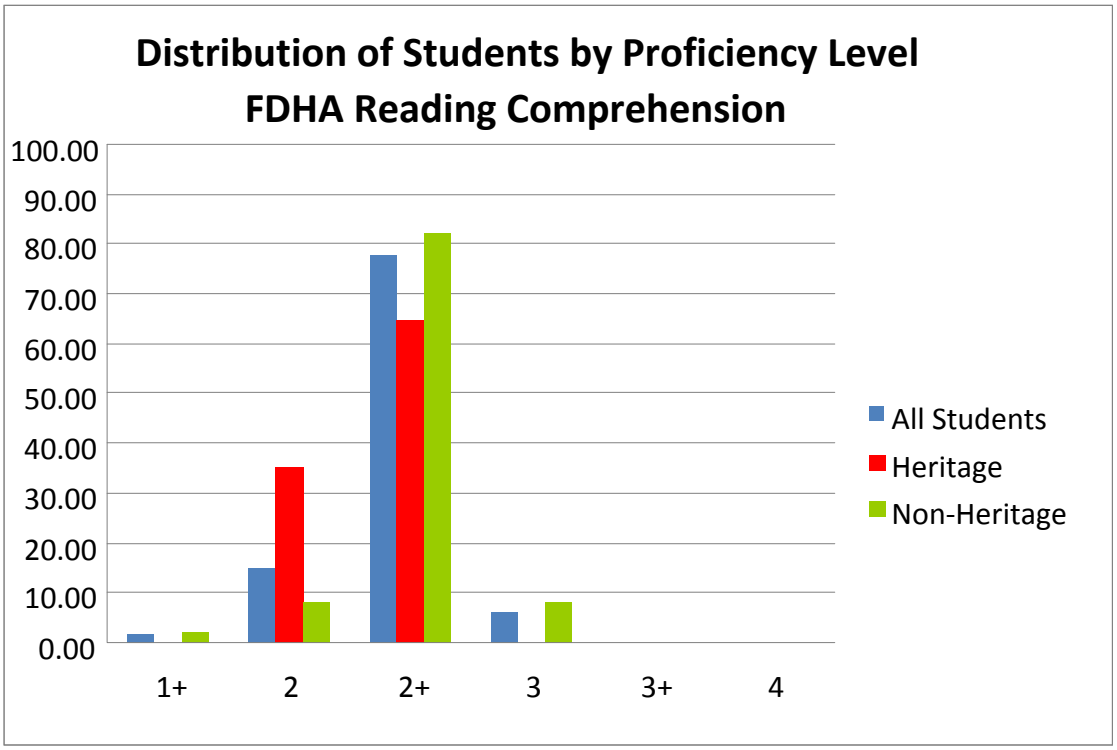
Table 7. *Proposed Cut Scores FDHA Russian Reading Comprehension: Round Two (Final)*

Panelist	Converted Item Measure for 1+/2	Converted Item Measure for 2/2+	Converted Item Measure for 2+/3	Converted Item Measure for 3/3+	Converted Item Measure for 3+/4
1	306	480	566	641	650
2	360	487	538	630	650
3	306	480	590	650	666
4	306	481	590	630	650
5	352	480	590	641	724
6	306	481	590	641	666
7	306	487	590	641	666
8	306	481	590	650	666
9	302	480	538	641	650
Mean	317	482	576	641	665
St. Dev.	22.43	2.93	22.81	7.09	23.41
Median	306	481	590	641	666
Mode	306	480	590	641	650

Table 8. *Final Cut Scores and Impact Data for FDHA Russian Reading Comprehension Section at RP67*

ILR Proficiency Level	Cut Score	Percent of Students	Cumulative Percent of Students
1+	386	1.49	1.49
2	551	14.93	16.42
2+	645	77.61	94.03
3	710	5.97	100.00
3+	734	0.00	100.00
4	---	0.00	100.00

Figure 4.
Distribution of Students in Percentages by ILR Proficiency Levels and by Heritage Information at RP67: FDHA Russian Reading Comprehension



REFERENCES

Brown, A. (2009). Less commonly taught language and commonly taught language students: A demographic and academic comparison. *Foreign Language Annals*, 42(3), 405-423.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: SAGE Publications.

Davidson, D. E. & Garas, N. (2009). *The ACTR Nationwide Survey of Russian Language Instruction in U.S. High Schools in 2009*, Russian Language Journal, Vol. 59, pp. 3-20.

Furman, N., Goldberg, D., & Lusin, N. (2007). *Enrollments in languages other than English in United States institutions of higher education, Fall 2006.* Retrieved December 30, 2009, from http://www.mla.org/pdf/06enrollmentsurvey_final.pdf

Janus, L. (2000). An overview of less commonly taught languages in the United States. *NASSP Bulletin, 84(612),* 25-29.