R-CAPE: A Russian Computerized Adaptive Placement Exam

Jerry Larson and Marshall R. Murray
BrighamYoung University

Introduction

During the past several years interest in studying Russian at Brigham Young University has increased dramatically while a number of high school programs across the country now list Russian as part of their language offerings. This increase in Russian enrollments and the wide disparity in high school programs has resulted in a less homogeneous group of students wishing to study Russian at BYU. A few years ago almost all students enrolling in their first college Russian course signed up for Russian 101 (Beginning Russian). Now, however, many students are past the beginning course stage and are prepared to commence their college Russian study in a second- or third-semester Russian course. This shift in pre-college preparation has created a need for placement measures for our Russian programs.

> *"During the past several years interest in studying Russian has increased dramatically; a number of high school programs across the country now list Russian as part of their language offerings."*

Because of increased communication and commerce with Russian-speaking countries, Russian language programs have grown significantly. Consequently, Russian language departments have struggled with placement decisions, trying to determine as accurately and efficiently as possible which courses are best suited for their incoming students. They have tried various methods for deciding where students should begin, including "seat time" (i.e., amount of classroom exposure to the language) decisions and paper-and-pencil placement tests. These procedures were less than satisfactory, however, since seat time is not a particularly good indicator of ability—given the great disparity in secondary school Russian programs, teachers, and students—and paper-and-pencil tests tend to be an administrative nuisance: they require a great deal of time to schedule, administer, and score, and then the individual students have to be notified of their results before they can complete their registration.

Given the need for a more efficient way to determine placement levels in Russian and having experienced successful placement of students in Spanish, French, and German via computer-administered placement exams, efforts were undertaken to develop a computerized adaptive placement test for Russian.

Computer Adaptive Testing

Though still a fairly recent innovation in language testing, computer adaptive tests have been recognized by language testing specialists as an extremely efficient and effective method of assessing language skills, particularly the receptive skills (Dandonoli 1989; Henning 1987; Henning 1991; Larson 1989; Larson & Madsen 1985). A considerable amount of interest of late has been devoted to exploring the feasibility of using computer adaptive tests for proficiency and diagnostic testing, as well as placement testing. (See Proceedings of the Seminar on Issues in Computer Adaptive Testing of Second Language Reading Proficiency sponsored by the Center for Advanced Research on Language Acquisition, University of Minnesota, Minneapolis, Minnesota, March 20-22, 1996, forthcoming.)

How Computer Adaptive Tests Function

For the information of those who may not be familiar with the concept of computerized-adaptive testing, we will briefly describe the theory behind it. The computer acts as the "examiner," evaluating each response of the examinee and then presents him or her with an "appropriate" subsequent item from the test item bank that resides on a floppy or the hard disk of the computer. Each item presented during the test depends upon the correctness of the examinee's response to the previous item, i.e., if an item is answered correctly, a more difficult item is presented; if the response is incorrect, an easier question will be given. Essential to the success of this testing procedure is the difficult calibration of the test questions in the item bank. Prior to final selection for the item bank, each item undergoes a a specialized (Rasch) statistical analysis in which its performance as a test time is evaluated and a difficulty index is assigned. The computer selects items during the administration of the test based on a calibrated scale of item difficulty formed by these indices.

Benefits of Computer Adaptive Testing

Due to the computer's unique capabilities and its central role in delivering the test, several benefits are realized. Computer adaptive tests can be, and generally are, "power" tests rather than "timed" assessment measures; thus, each student takes the exam independently and can work at his or her own pace, which tends to lessen the anxiety of having to hurry to finish a test. Additionally, since computerized exams are given independently, they allow for impromptu testing: whenever a student is ready or available, he or she can be tested without having to wait for a roomful of students to be assembled on a given day before the semester begins, as is often the case with traditional placement testing.

Because the computer is able to keep track of the examinee's performance as the test progresses, a detailed or simplified test report can be generated immediately upon completion of each test and displayed on the screen (or printed for archival purposes), making it possible for students to receive timely feedback regarding their test performance. In the case of placement testing, this intantaneous reporting capability is very important, since it precludes students' having to wait several hours–or sometimes days–before receiving their test results so they will know in which course they should enroll.

"Computer adaptive language exams provide a "common yardstick" measurement. That is, all examinees who take the test are evaluated on the same ability/ difficulty scale. This allows accurate comparison of students from one semester to another, or even from one school to another, without concern for which text was used, who the teacher was, or any other external conditions pertaining to the students' learning."

One of the great advantages of a computer-delivered test is that many administrative costs and time-consuming procedures are eliminated or greatly reduced. For example, it is not necessary to pay a test administrator to gather several students together, pass out test forms, answer sheets, pencils, etc., collect them, then score the tests and contact the students regarding their results and proper placement. And since the entire test is contained on a single three-and-a-half-inch floppy disk or a computer's hard disk, storage requirements and test security concerns are greatly reduced. Departments don't have to occupy filing cabinets with alternative test forms, and the items on the disk are compiled in such a way as to be inaccessible to anyone who tries to read them.

Besides the benefits possible through using the computer as the test "administrator," there are additional advantages due to the adaptive nature of the test. Because the exam uses a pool of previously calibrated test items, a computer adaptive exam provides multiple equated test forms, meaning that each examinee receives a unique, yet equivalent, version of the test. It is also virtually impossible for a student to cheat during the exam by looking at a neighboring examinee's computer screen, since the computer selects items from level to level from a pool of several items at the various difficulty levels; therefore, the chance of two examinees sitting next to each other receiving the same test items is very unlikely.

Computer adaptive language exams provide a "common yardstick" measurement. That is, all examinees who take the test are evaluated on the same ability/difficulty scale. This allows accurate comparison of students from one semester to another, or even from one school to another, without concern for which text was used, who the teacher was, or any other external conditions pertaining to the students' learning.

The capability of the test to adapt to the ability level of the examinee yields other benefits as well. For example, test efficiency is improved. Students no longer have to waste time on items that are much too easy or much too difficult. Such items give no valuable discrimination information anyway; they simply tend to bore or frustrate the examinee. Hence, an adaptive test greatly improves examinee attitude. Follow-up studies we have conducted verify this fact: students surveyed greatly preferred the computer adaptive tests to the traditional paper-and-pencil placement tests (Larson, 1989).

Limitations of Computer Adaptive Testing

While computer adaptive tests present a number of advantages over other more conventional methods of testing, they do have some limitations, some of which can be reduced to a certain extent. One of the most common objections that has been levied against computerized testing is the expense involved. While it is true that computers are expensive, it is also true that nearly every language department has access either to computers in a lab on campus or to various office work stations. A small number of computers (two or three) is sufficient to administer forty to fifty tests during normal working hours.

Another limitation is the amount of text that can be presented on the computer screen at one time. The size of a reading item, for example, is restricted to the size of the computer screen. Given that the size of the font needs to be large enough for the examinee to read the item easily, the maximum amount of lines of text for each reading passage is about twelve to fifteen, leaving space for the question stem and answer options. Particularly for higher-ability level students, this presents a problem, since it would generally be preferable to have longer reading passages. However, this limitation can be mitigated by programming the test delivery shell to allow the examinee to scroll the items when necessary. This is fairly easily done now with scrolling fields in the Windows environment on the PC and with the standard Macintosh operating system.

"Until answer-judging routines via artificial intelligence become more refined, we are basically limited to testing receptive skills only, meaning that acceptable computerized tests of speaking and writing are not yet possible."

Computer-delivered tests are definitely restricted at this point in time with respect to the kinds of items they can evaluate. Until answer-judging routines via artificial intelligence become more refined, we are basically limited to testing receptive skills only, meaning that acceptable computerized tests of speaking and writing are not yet possible.

Another criticism levied against computer-delivered tests is the assumption that they cause too much test anxiety. Studies conducted at Brigham Young University (Larson, 1989) do not corroborate this assumption. In fact, we found just the opposite. Students who had little or no previous experience with computers indicated that taking the placement test via the computer did not cause them more anxiety than taking a traditional paper-and-pencil version. In fact, students with absolutely no previous computer experience were the ones who showed the greatest preference for the computerized version of the test.

R-CAPE Purpose and Target Population

Since 1986, Brigham Young University has developed three computer adaptive placement exams: Spanish (1986), French (1989), and German (1990). The tests were created to assist in the initial placement of students into college-level language courses. In 1994, to meet the need for a placement measure for Russian at BYU and in response to requests from users of previous CAPES, BYU began a two-year development and testing project which resulted in the Russian CAPE (R-CAPE). Again, the test was developed specifically to place new or transfer college students who have previously been exposed to Russian through classroom or life experience. As with the previous CAPES produced by BYU, the R-CAPE is designed to place students in the first, second, third, or higher semester Russian courses.

Development of the R-CAPE

The two-year process of developing the R-CAPE can be divided into five phases which were accomplished sequentially:

1) Test item creation,

2) Item analysis, calibration, and selection,

3) Integration of the test bank into the computer adaptive engine,

4) Determination of cutoff scores for each placement level, and

5) Validity and reliability assessment.

Test Item Creation: First, a table of specifications was developed to describe the cognitive domains to be included in the R-CAPE and the type of test items that would assess an examinee's knowledge of these domains. To maintain consistency with previous CAPEs developed by Brigham Young University, the R-CAPE continued to assess the three language domains—grammar, vocabulary and reading. The specific competencies assessed by R-CAPE for each of these domains are: 1) grammar—the ability to select correct grammatical forms and endings for the appropriate syntactic environment, 2) vocabulary—the ability to express knowledge of Russian words and expressions, 3) reading—the ability to read and understand written passages in Russian.

A student's knowledge of grammar, i.e., the structure of the language, vocabulary, i.e., the lexis of words with which to understand the language, and reading, i.e., the integration of grammar and vocabulary, all reflect previous exposure and experience with the language and are appropriate measures of ability at various levels of college-level Russian language courses. In the initial stage of R-CAPE development, a large number of multiple-choice format test items were created, from which eventually a pool of approximately 450 serve as the item bank used by the R-CAPE computer program. To ensure that 450 test items survive the scrutiny of Item Response Theory (IRT) performance analysis, 1200 test items (400 each for grammar, vocabulary and reading) were created at the outset. These were further subdivided into three relative levels of difficulty to test examinees of varying background and experience in Russian. These items of varying difficulty were based on grammatical concepts and vocabulary presented in several beginning and intermediate college-level texts. Reading items were obtained from non-copyrighted authentic Russian materials taken from a variety of literary formats (e.g. dialogue, poetry, narrative, etc.) In this process, an initial attempt was made to divide items of each domain into three approximate difficulty levels. This initial item leveling procedure underwent a verification process at two levels. First, a native-speaking BYU Russian language instructor and a BYU Russian language professor screened all test items for accuracy and relative level of difficulty. The second verification of test item difficulty levels occurred during a statistical item analysis, which occurred in a later phase of the CAT development process.

"A student's knowledge of grammar, i.e., the structure of the language, vocabulary, i.e., the lexis of words with which to understand the language, and reading, i.e., the integration of grammar and vocabulary, all reflect previous exposure and experience with the language and are appropriate measures of ability at various levels of college-level Russian language

Item analysis, calibration and selection: The most complex and certainly one of the most critical phases of the R-CAPE development process was the calibration and selection of test items for the test bank. The approximately twelve hundred test items were divided equally into eight test forms of one hundred fifty items. The forms were linked together with thirty anchor test items to allow for creation of a common scale of item difficulty for the entire set of test items. The one hundred ninety-four students who participated in test item analysis consisted of volunteers from five U.S. colleges and universities: Brigham Young University, the Ohio State University, Ricks College, the University of California at Berkeley, and the University of Texas at Austin. These students provided a diverse group of examinees to evaluate the performance of each test item and to calibrate the items on a common difficulty scale.

The performance of potential test items was analyzed using Bigsteps™, a Rasch one-parameter item response theory (IRT) computer program.[1] Each of the 194 examinee's responses to a 150-item form of the test was entered into the Bigsteps™ program and underwent twelve IRT analysis iterations. After each iteration, items were removed that did not meet a requisite level of performance. This level of test item performance required that an examinee or test item perform in a manner consistent with the computer algorithm of expected performance as evidenced by each examinee's history of responses on the calibration test. The IRT analysis process reduced the pool of candidate test items from 1200 to approximately 700. The resulting pool of qualified items was reduced further by selecting as even a distribution as possible of items by skill area (grammar, vocabulary, reading) for each of approximately 50 calibrated levels of difficulty. The remaining approximately 225 well-performing and calibrated test items that were not selected were retained for potential future use in the R-CAPE, if needed.

Integration of items into the test bank: During the calibration procedure, each of the test items was assigned a difficulty index number. The chosen items identified by their respective index number were placed into the R-CAPE test item bank. Using these index numbers, the testing computer program is able to branch from easier to more difficult items, and vice-versa, as required by an examinee as he or she progresses through the test.

Testing methodology: The R-CAPE test uses the same testing shell as the other CAPE tests. This shell obtains student demographic data before the test begins, creating an individual data file for each examinee. The computer algorithms within the shell also give preliminary instructions and a sample test item to the examinees, guiding them through the testing process. Once the test proper begins, the computer calculates an ability estimate of the examinee and selects a subsequent item based on that person's demonstrated ability. This is possible by matching the examinee's ability estimate with an appropriate item difficulty coefficient. As this item selection procedure continues, the computer also calculates the current error of measurement. Once this error is reduced to less than .40 through several item administrations, the test terminates. Upon completion of the test, the computer displays the examinee's score on the screen, thus providing instantaneous performance feedback to the student. The student may then use this score to determine in which course he or she should enroll.

Determination of cutoff scores: The scores that correspond to each course in question (i.e., the "cutoff" scores) are determined by administering the test to a number of students at each course level, e.g., Russian 101, 102, 201. By plotting the students' scores by course, it is possible to determine the range of scores that is acceptable for those courses in an individual institution's curriculum.

"By plotting the students' scores by course, it is possible to determine the range of scores that is acceptable for those courses in an individual institution's curriculum."

R-CAPE Delivery System

As mentioned above, the same computer algorithm used to administer the other CAPE tests is incorporated in R-CAPE. Initially, the CAPE tests were created for the DOS environment. The R-CAPE, however, operates in the MS Windows (3.1 or higher) environment and requires an Intel 386 or higher computer processor with a minimum of 4K RAM (at least 8K is recommended).

Test Analysis

Validity. Content validity was a key consideration in the initial stages of test item development. At the outset, items of varying difficulty were based on grammatical concepts and vocabulary presented in several beginning and intermediate college-level texts. Next, prospective test items were subjected to the scrutiny of two Russian language experts who were teaching Russian in BYU's Department of Germanic and Slavic Languages. Both experts exercised the right to disqualify or to require modification of test items that were not representative of the initially assigned difficulty level, had faulty or multiple correct distractors, or in some way were in error. An additional attempt to build evidence of content validity occurred at the completion of the R-CAPE item analysis and calibration process. The course grades of 56 Russian language students at the end of the semester were compared with each student's score on the final version of the R-CAPE. The correlation coefficient of students' performance between course grades and R-CAPE scores was .86.

Reliability. Measures of internal reliability of the test items undergoing item response theory analysis was provided after each IRT analysis iteration by the Bigsteps program. Reliability coefficients for the twelve IRT analysis iterations ranged from .76 to .80. At the completion of item development, a test-retest reliability assessment was conducted using Russian language students at Brigham Young University. The final R-CAPE version was administered to approximately 140 students of Russian at the end of their fall semester 1995. Two weeks after the initial testing, a retest was administered to 33 volunteers, representing a variety of Russian language levels. The reliability coefficient of the test-retest reliability was .96.

Conclusion

Though there are some limitations and disadvantages associated with computer-adaptive language testing, this method of placement assessment for foreign languages has proven to be an efficient and effective means of placing entering language students into an appropriate initial course of college-level language study. In addition, computer adaptive testing may represent a significant savings of time, energy—and even money—over administering comparable paper-and-pencil tests.

At Brigham Young University, the Russian-CAPE, like its predecessors for Spanish, French and German, is already being used in the placement process and has become a significant aid for helping Russian students enroll in courses appropriate to their ability level.

Additional research is required to refine the computer adaptive placement process by comparing placement test scores with performance at a variety of course levels. Further work is also needed to investigate the use of other item formats in computer adaptive tests.◆

Notes

[1] Bigsteps™ is available through MESA Press, 5835 S. Kimbark Ave., Chicago, IL 60637-1609.

Works Cited

Dandonoli, P. 1989. "The ACTFL Computerized Adaptive Test of Foreign Language Reading Proficiency," 291-300 in W. F. Smith, ed., Modern Technology in Foreign Language Education: Applications and Projects. Skokie, IL: National Textbook Company.Henning, G. T. 1987. A Guide to Language Testing: Development, Evaluation, Research. Cambridge,

_____.1991. Validating an Item Bank in a Computer-assisted or Computer-adaptive Test: Using Item Response Theory for the Process of Validating CATS. Chapter 10 in Computer-assisted Language Learning and Testing: Research Issues and Practice, edited by P. Dunkel. New York: Newbury House.

Larson, Jerry W. 1989. "S-CAPE: A Spanish Computerized Adaptive Placement Exam," 277-291 in W. F. Smith, ed., Modern Technology in Foreign Language Education: Applications and Projects. Skokie, IL: National Textbook Company,.

Larson, J. W., and Madsen, H. S.1985. "Computerized Adaptive Language Testing: Moving beyond Computer Assisted Testing," CALICO Journal, 2, 3:32_36.

Jerry W Larson, Ph.D., FL Education, University of Minnesota; M.A., Spanish Pedagogy, Brigham Young University. Currently Director of the Humanities Research Center and Professor of Spanish at Brigham Young University. Marshall R. Murray, M.A. in Language Acquisition with an emphasis in Russian, Brigham Young University. Presently pursuing a Ph.D. in Instructional Science at BYU. He currently serves as the Director of the English Language Study Center in Salt Lake City, Utah.