VELC Test for testing competency: Verification of reliability and validity

Tetsuhito Shizuka, Daito Bunka University
Masamichi Mochizuki, Reitaku University

Abstract

The authors have developed a new competency test to make visible the English-language skills of Japanese university students as much as possible. The test divides the two sections of listening and reading into three parts each, measuring listening vocabulary (L1), connected speech deciphering (L2), and listening comprehension (L3) along with reading vocabulary (R1), sentence structure awareness (R2), and reading comprehension (R3). Equating data from trial testing of approximately 5000 Japanese university students, using a Rasch model, makes it possible to compare scores on the same scale no matter which of multiple forms the test takers used. The test's coefficient of reliability is higher than 0.95, and its multiple correlation coefficient to TOEIC scores is 0.82. Feedback on results is provided through a Web-based e-Portfolio that can be described as a record of an individual's English-language ability. Students also can use this test to ascertain changes in their own English-language abilities by taking the test periodically. As a result, it can be expected to see a variety of uses that have not been possible with previous one-time testing.

I. Development objectives

One could say that until now there has been no test of competency that provided swift, detailed feedback on the results of efficiently measuring the English-language skills of Japanese university students. Drawbacks of traditional tests have included content not suited to university students, unsuitable degrees of difficulty, and slow feedback on results. Recognizing this situation, we began development of a new test of competency intended to overcome these drawbacks.

Since the goal was to develop a test for placement, measuring results of learning, and diagnosing weak points in ways suited to the actual abilities of Japanese university students, we named the test the Visualizing English Language Competency (VELC) Test. Intending it for use particularly in placement of new students and measurement of changes over time in the English-language abilities of existing students, we began designing the test with the physical requirements that it be able to be conducted with time to spare within a 90 minute class session in a normal university classroom and that feedback on results be available within a few days.

II. Specifications of the test

Based on the above requirements, we first decided that the test should measure, through multiple-choice questions, students' listening and reading abilities, which among the four basic language skills are those related to receptiveness, and then we carried out repeated study of the subskills related to these two skills that could be considered particularly important for Japanese university students. In the end, we decided on the test Specifications shown in Table 1.

Table 1. Specifications of the VELC Test

Part*	Question format
Listening 1	Listening to a word in Japanese and then choosing from four choices presented aurally the corresponding English word.
Listening 2	Listening to a short English sentence and then choosing from four choices presented visually the word in the designated position.
Listening 3	Listening to an English passage of a certain length, guessing the word replaced by a beep, and choosing it from four choices presented visually.
Reading 1	Viewing a word in Japanese and then choosing from four choices presented visually the corresponding English word.
Reading 2	Reading a sentence with one word missing and then choosing from four choices the position in the text where the designated word should be placed to make it a complete sentence.
Reading 3	Choosing from four choices the words that should fill in the blanks in an English passage of a certain length (roughly 30-80 words).

* Note: Each part consists of 20 questions, for a total of 120 questions.

Part 1 of each of the sections on listening and reading measures the constructs of size of listening vocabulary and size of reading vocabulary, respectively. Target words were chosen broadly from JACET 8000 levels 1-7, and the question formats were based on the Mochizuki Test (Mochizuki, 1998). Listening part 2 uses a multiple-choice format for partial dictation, intended to measure the ability to separate a continuum of speech sounds into meaningful segments. Listening part 3 is a type of cloze test (Oller, 1979) for listening, which could be considered capable of measurement of overall listening ability. Reading part 2 is a type of invisible-gap filling test (Shizuka, 2004) to measure the ability to analyze correctly the structure of a relatively long sentence. While reading part 3 is, like its listening counterpart, a type of cloze test, the multiple choices have been designed to require an understanding of the broader context in order to fill in the blanks.

III. Development process

With the exception of part 1, which used a vocabulary list, all of the English materials used in the test were written by native speakers belonging to the study group. This was conducted by considering the English-language abilities needed by Japanese university students, employing a policy of focusing chiefly on academic content concerning scientific and social subjects. After a process of studying and revising within the study group the English materials converted into questions, a total of three rounds of trial problem-solving was conducted with the participation of more than 5000 Japanese university students, as outlined below.

A. First trial round

Each of multiple groups of participants was tested using prospective questions in one to three types of formats, the results were analyzed using Winsteps (Linacre, 2005), software for Rasch modeling, and questions that fit the model were chosen, looking overall at indicators such as Infit Mean Square.

B. Second trial round

Groups of questions with good compatibility based on the results of the first trial round and that also were judged to have stable degrees of difficulty were chosen as links, and new groups of participants were given tests that added unused questions to these, to increase further the number of quality questions. From the results of the first and second rounds of trials, at a minimum all questions within the same part were placed on a scale of difficulty, excluding incompatible ones.

C. Setting up forms

Multiple forms (covering all six parts for listening and reading) of 120 questions each, designed so that all forms would have largely identical degrees of difficulty, were set up using tentatively finalized degree of difficulty values for the questions.

D. Third trial round

New groups of participants were tested using the multiple forms set up, to further collect data on problem-solving.

E. Deciding on final form

All problem-solving data from the total of three trial rounds were used to recalculate the degrees of difficulty of the questions, and the final multiple equated forms were decided on based on the resulting figures.

IV. Results from the VELC Test provided as feedback

This test provides feedback in the forms of three main types of information.

A. VELC Score

In this test, six types of VELC scores are calculated: the general score, the listening score, the reading score, the listening vocabulary score, the connected-speech deciphering score, the listening comprehension score, the reading vocabulary score, the sentence-structure awareness score, and the reading comprehension score. To facilitate interpretation of the results, each type of VELC score is scaled such that the average score of the Japanese university students who took the trial tests was 500 and the standard deviation 100.

For example, if a future taker of this test gets a listening score of 550 and a reading score of 450, then he or she can be considered to have listening ability 0.5 times the standard deviation higher than the average among Japanese university students (in the 31% top percentile) but reading ability 0.5 times the standard deviation lower than the average (in the 31% bottom percentile). (See Table 2.)

Table 2. VELC score percentile rankings

Velc Score	Bottom percentile rank	Top percentile rank
250	1%	99%
300	2%	98%
350	7%	93%
400	16%	84%
425	23%	77%
450	31%	69%
475	40%	60%
500	**50%**	**50%**
525	60%	40%
550	69%	31%
575	77%	23%
600	84%	16%
650	93%	7%
700	98%	2%
750	99%	1%

In calculating these VELC scores, we used the Winsteps UPMEAN and SCORE FILE commands.

B. Breakdown by knowledge and skill

While the above VELC scores roughly show norm-referenced skill profiles by skill areas, the "breakdown by knowledge and skill" provides feedback on more detailed types of subknowledge and subskills. For this purpose, we grouped all 120 questions included on each form into the categories shown in Table 3. These categories were identified through discussion among multiple research group members. Categorization of questions was conducted independently by multiple evaluators and completed through discussion to reconcile their results. Based on the final category table, we calculated the average percentage of correct answers for all participants on questions belonging to each category, for each form. VLEC Test takers can confirm their own rankings on these subknowledge and subskill categories by comparing their own percentages of correct answers in each category to national averages.

Table 3. Standards for categorizing questions by knowledge and skill

Category	Type(s) of questions in category
High-school vocabulary	Correct answer is a JACET8000 level 1-2 word
Basic university vocabulary	Correct answer is a JACET8000 level 3-4 word
Applied university vocabulary	Correct answer is a JACET8000 level 5-7 word
Content word comprehension	Aural question for which the correct answer is a content word
Weak forms of function words	Aural question for which the correct answer or immediately preceding word is a weak form of a function word
Unreleased stop	Aural question for which the correct answer or immediately preceding word has an unreleased stop
Ambiguity	Aural question for which the correct answer and the words before and after it include an ambiguous vowel schwa
Linking	Aural question for which the correct answer and the immediately preceding word are linked in the form C+V
Phoneme identification	Aural question with a distractor phoneme similar to the correct answer
Long subject	Question including a subject consisting of five or more words
Long object	Question including an object consisting of five or more words
Long prepositional phrase	Question including a prepositional phrase consisting of five or more words
Long adverbial clause	Question including an adverbial clause consisting of five or more words
Relative clause	Question including a relative clause
Postmodifier/ description	Question including a postmodifier consisting of five or more words
Distant agreement	Question including an agreement relationship between two or more sentences or separated by five or more words.
Combination of elements	Question including five or more noun phrases or verb phrases connected by and or by or
Insertion into sentence	Question including an inserted expression
Appending to end of sentence	Question including an expression appended to the end of a sentence
Relation between sentences	Question requiring understanding of the relation between sentences

C. "Can do" level by situation

While both VELC scores and breakdown by knowledge and skill provide norm-referenced information, it is the test taker's "can do" level by situation, which provides criterion-referenced information, that shows what he or she is likely to be capable of accomplishing in English, and the degree of such accomplishment.

First, we collected answers from approximately 550 test takers in the second trial round to questions on what they considered to be their own degrees of understanding for the 10 categories of listening situation and 10 categories of reading situation shown in Table 4, using a scale of five grades 0-4, with 0 indicating an understanding of 0-10% of the content and 4 indicating an understanding of 90-100%. Rasch modeling of the data on these answers with the data on solving the listening and reading questions can be used to calculate the successful degree of difficulty (D) in each situation, as a logit value. Substituting this logit value and the ability logit value (B) of each test taker into the formula $Pr = \exp(B-D)/(1+\exp(B-D))$ gives the test taker's probability of success in that situation. Table 3 shows the probability of success in each situation of a participant in this trial who demonstrated average ability (B = 0.0), rounded off to the nearest 5%.

Table 3. "Can do" level of an average university student, by situation

Listening to a recording read from middle-school level teaching materials	80%
Listening to simple English instructions provided by a Japanese instructor during class	75%
Listening to a recording read from middle-school level teaching materials	50%
Taking an English course taught by a native-speaking instructor	40%
Listening to an English song with a slow tempo (such as a ballad)	30%
Listening to departure and platform announcements in English at airports, railway stations, and elsewhere while traveling overseas	25%
Listening to the answers of staff to your questions in a restaurant or shop while traveling overseas	25%
Watching a news program produced overseas	10%
Watching an English motion picture without subtitles	10%
Listening to native speakers speaking naturally to each other	5%

Reading an English text written for use as middle-school level teaching materials	90%
Reading an English text written for use as high-school level teaching materials	65%
Reading English signs (such as those showing how to buy tickets and fares) in public transportation stations overseas	50%
Reading the text of a simple English email addressed to you	50%
Reading the definitions in an English (not bilingual) dictionary (such as Longman) edited for learners	35%
Reading the names and descriptions of dishes on an English menu at a restaurant overseas	35%
Reading a short novel rewritten for learners	25%
Reading domestic news in an English newspaper published in Japan	25%
Reading the news in an English newspaper published overseas	10%
Reading a best-selling English novel published overseas	10%

V. Reliability and validity

(1) Reliability

For test scores to be reliable, first of all the degree of difficulty of the questions must be appropriate for the test taker's level. If the questions are either too easy or too difficult the amount of information produced by the results will be small. Fig. 1 shows the distribution (at left; each pound sign [#] represents 15 persons) of the 5583 participants tested on the questions ultimately employed in Form A (tentative title) and ability/difficulty of 120 questions (at right; each " X" represents one question). The questions are spread almost uniformly across a broad range of 4.5 logits from -3 logits to roughly 1.5 logits, a visual indication of the fact that the test is suitable for a wide range of test takers from lower to higher levels and has a degree of difficulty that is appropriate for Japanese university students overall.

Next, we calculated coefficients of reliability for the data on the 226 persons who answered at least 115 of the 120 questions included in Form A. The results showed that the Rasch person reliability, corresponding to raw-score based Cronbach's alpha, was .95 while the Rasch item reliability also was .95. Thus, the test can be said to show a very high level of reliability for this group of test takers.

(2) Criterion-related validity

Next we will look at the results of multiple regression analysis with the VELC score as the predictor variable and the TOEIC score as the target variable, as data on criterion-related validity. The subjects of this analysis were the population (N=375) of trial participants who provided all three of their TOEIC scores: their listening, reading, and total scores.

The models ultimately chosen through a stepwise multiple regression analysis in which the target variables were TOEIC listening and reading scores and the predictor variables were listening vocabulary (L1), connected-speech deciphering (L2), listening comprehension (L3), reading vocabulary (R1), sentence-structure awareness (R2), and reading comprehension (R3) are shown below.

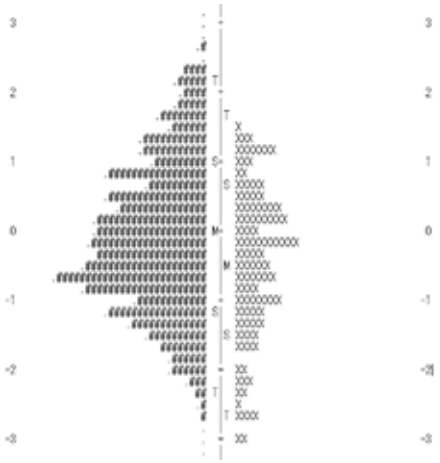

Fig. 1. Distribution of test taker ability and question difficulty on Form A

$$\text{TOEIC L} = -74.886 + 0.075*L1 + 0.199*L2 + 0.248*L3 + 0.119*R3$$

$$\text{TOEIC R} = -199.599 + 0.075*L1 + 0.079*L2 + 0.148*L3 + 0.109*R1 + 0.174*R2 + 0.211*R3$$

The coefficients of determination for these models were 58% and 64%, respectively. In addition, prediction of total TOEIC scores by totaling the predicted values from these models resulted in a coefficient of determination of 68%, which corresponds to a multiple correlation coefficient of 0.82. Fig. 2 shows a plot of predicted values and actual measurements. It is quite interesting that this 120-question test, which could be completed in 70 minutes, has a high correlation of 0.82 to results of the TOEIC exam, which takes two hours. The coefficient of determination of 68% can be said to indicate predictive precision suitable for practical use.

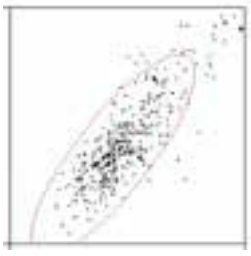

Fig. 2. Predicted (X axis) and actual (Y axis) TOEIC scores

(3) Construct validity

Next, we attempted a factor analysis for eight variables: the six parts of the VELC Test (L1, L2, L3, R1, R2, and R3) and the two sections of the TOEIC test (L and R). Fig. 3 shows the model built

using Amos based on the loadings of three factors identified as a result of exploratory factor analysis using SPSS. As expected from our theoretical analysis, the model in which VELC parts L2 and L3 and the TOEIC listening section form one factor (listening), VELC parts L1 and R1 form another factor (vocabulary), and VELC parts R2 and R3 and the TOEIC reading section form a third factor (reading) demonstrated high validity (GFI = .957, AGFI = .908).

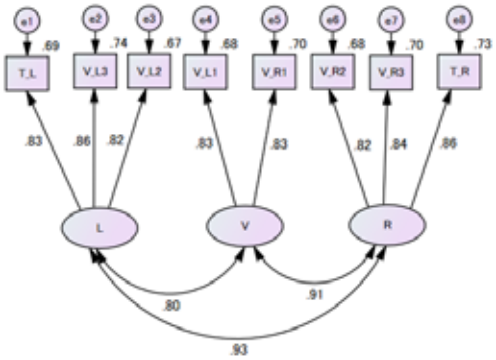

Fig. 3. Factor structure of scores on the six parts of the VELC Test and TOEIC section scores

(4) 2012 testing data

Next we will look at data on the results of 2327 students from 19 universities who took the Form A test in 2012.

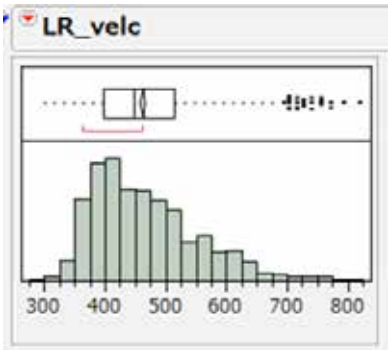

Fig. 3. Distribution of scores of Form A test takers (N = 2327)

Fig. 3 shows a histogram of the scores of all test takers. Scores are distributed across a broad range from the lowest score of 293 to the highest of 820, with an average score of 462. The distribution of this sample shows a fairly long tail on the right-hand side. From this histogram we can see that VELC Test scores are distributed across a wide range.

Looking at the question of whether scores vary by university, Fig. 4 shows a box plot of scores by university.

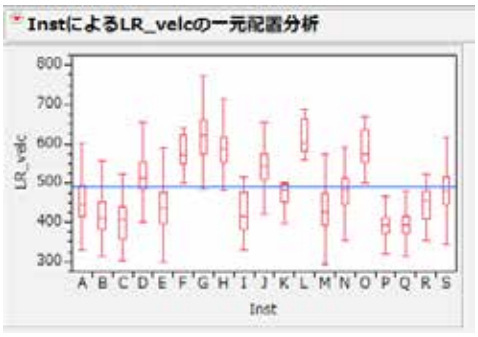

Fig. 4. Distribution of scores of 19 universities where students took the Form A test

It is clear at a glance that the distribution of scores varies substantially by university. University G had the highest average score (M = 623), while University Q had the lowest (M = 396). An analysis of variance showed that overall there was a significant difference ($F(18, 2308) = 204.02$, $p < .000$). These can be interpreted as showing that the VELC Test clearly reveals differences in English ability among universities.

VI. Conclusion

The VELC Test has been developed as a test of competency suited to the actual conditions of Japanese university students. Its degree of difficulty matches the average level of Japanese university students and its coefficient of reliability is very high. It also has a high degree of criterion-related validity using TOEIC scores as the target criterion, predicting 64% of the TOEIC score variance. Its distribution of scores also can be considered to reflect levels of English ability faithfully when testing groups of students with different ability levels.

Since the test can be completed in 70 minutes, it could be conducted multiple times in a year without having a major impact on courses. Since data on test results can be checked on the Web, it can enable students to check the growth in their English abilities in detail by taking the test on a continual basis. It can be said to be suited to a wide range of uses, including assigning students to classes according to their competency levels and measuring course results.

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model.* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Linacre, J. M. (2005). Winsteps (Version 3.55) [Computer software]. http://www.winsteps.com/

Mochizuki, Masamichi. 1998b "Nihonjin eigo gakushusha no tame no goi saizu tesuto [A vocabulary size test for Japanese learners of English]." *The IRLT Bulletin, No. 12*, pp. 27-53.

Oller, J. W., Jr. (1979). *Language tests at school.* London: Longman.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmark Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago.)

Shizuka, T. (2004). Reliability and validity of "invisible gap filling"items. *JLTA Journal, 6*, 108-127.