An Experimental Report of the Web-based Test for English Communication (WeTEC) and the Common European Framework of References (CEFR)

Yasuko Nogami¹, Michiko Nakano², Hiroshi Matsuda¹, Norio Hayashi¹

¹The Japan Institute for Educational Measurement, Inc., ²Waseda University

nakanom@waseda.jp

Abstract

The Web-based Test for English Communication (WeTEC) is used as a placement test as well as an achievement test at Waseda University. The WeTEC is a computer adaptive test and it has four sections: vocabulary, conversational schema, listening comprehension and partial dictation. Every year about 10000 students take this test and most of the test-takers study in the English Tutorials courses at the Open Education Center. English Tutorials are designed based on the Common European Framework of References (CEFR). There are six class levels: beginner, basic, pre-intermediate, intermediate, pre-advanced and advanced. These class levels roughly correspond to the six CEFR levels: A1, A2, B1, B2, C1 and C2. Since the WeTEC is also used as an achievement test, it has been hoped that correspondence between the WeTEC scores and the CEFR levels can be obtained. In this paper, we give our tentative solution to our problem.

Several kinds of the CEFR assessment questionnaires were given to both students and tutors. These questionnaires contain can-do descriptors. The analysis based on the Item Response Theory has demonstrated that both students and tutors judge the difficulty of descriptors in accordance to the six levels in the CEFR. However, the relationship between the WeTEC scores and the CEFR proficiency estimates was not strong enough to link the WeTEC scores and the CEFR levels. The present paper therefore attempts to link the WeTEC scores and the CEFR levels, analyzing the data by logistic regression.

Introduction

The English Tutorial Lessons at Waseda University have been offered by the Open Education Center since 1997, and they have been popular among students in the university. In the past four years, about 10,000 students took these Tutorial lessons per year. These courses aim to promote practical English skills so that the students can communicate functionally well in English. In order to achieve this objective, one tutor teaches a small group of four students. This small group training is effective to reduce students’ speech anxiety in English and to provide social contexts of speech situations. Since the tutorial lessons create a context for socialization, it can promote acquisition of English communicative competence and it is effective to let learners use their passive knowledge of vocabulary and grammar automatically and stably. The English Tutorials are designed based on the Common European Framework of References (CEFR) and there are six levels: beginners, basic, pre-intermediate, intermediate, pre-advanced and advanced. These levels roughly correspond to the six levels in the CEFR: A1, A2, B1, B2, C1 and C2 respectively. The Web-based Test for English Communication (WeTEC), a computerized adaptive test, has been used as a placement test to group students into the six levels.

Since the WeTEC is also used as an achievement test, it has been hoped that correspondences between the WeTEC scores and the CEFR can be obtained. In this research, several kinds of the CEFR assessment questionnaires were given to both students and tutors. These questionnaires contain can-do descriptors. The analysis based on the Item Response Theory (IRT) has demonstrated that both students and tutors judge the difficulty of descriptors in accordance to the six
levels in the CEFR. However, the relationship between the WeTEC scores and the CEFR proficiency estimates was not strong enough to link the WeTEC scores and the CEFR levels (Tsutsui, Nakano and Kondo, 2007). This paper therefore attempts to link the WeTEC scores and the CEFR levels, analyzing the data by logistic regression.

1 The Common European Framework of References (CEFR)

The CEFR was publicized in 2001 at the International Conference to Commemorate European Year of Languages held at the Free University, Germany, after a long-term empirical investigation in order to set up common standards in language teaching and learning. Since then, many books and articles about the CEFR have been published. It is now known as the framework providing the most comprehensive descriptors of language learning. The CEFR managed to define L2 proficiency in functional terms so that the same descriptors can be used to define learning goals, to develop learning materials and activities, and to judge learning achievement. At Waseda University, Michiko Nakano, the Director of English Language Division, Open Education Center, introduced CEFR into the framework of language teaching at Open Education Center in 2001. Nakano and her associates have modified teaching materials and tasks every year so that all the language learning activities in English Tutorials follow the CEFR descriptors and guidelines. In 2006 we started to validate the linkage of the tutorials to the CEFR. The present report is one of our efforts in this direction.

Table 1 gives an overall description of the six levels specified by the CEFR and Table 2 gives an example of the descriptors for phonological control.

Table 1: Overall description of the CEFR six levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Overall Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>Conveys finer shades of meaning precisely and naturally.</td>
</tr>
<tr>
<td>C1</td>
<td>Shows fluent, spontaneous expression in clear, well-structured speech.</td>
</tr>
<tr>
<td>B2</td>
<td>Express points of view without noticeable strain.</td>
</tr>
<tr>
<td>B1</td>
<td>Relates comprehensibility the main points he/she wants to make.</td>
</tr>
<tr>
<td>A2</td>
<td>Relates basic information on, e.g. work, family, free time etc.</td>
</tr>
<tr>
<td>A1</td>
<td>Makes simple statements on personal details and very familiar topics.</td>
</tr>
</tbody>
</table>

Table 2: CEFR descriptors for phonological control

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>As in C1</td>
</tr>
<tr>
<td>C1</td>
<td>Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.</td>
</tr>
<tr>
<td>B2</td>
<td>Has acquired a clear, natural, pronunciation and intonation.</td>
</tr>
<tr>
<td>B1</td>
<td>Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.</td>
</tr>
<tr>
<td>A2</td>
<td>Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time.</td>
</tr>
<tr>
<td>A1</td>
<td>Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/her language group.</td>
</tr>
</tbody>
</table>

2 WeTEC

The WeTEC is used as a placement test as well as an achievement test for the students who take English Tutorial lessons. It has been developed as a sister product of the Computerized Assessment System for English Communication (CASEC), which is one of the most widely-used computerized adaptive tests (CATs) in Japan (Nogami & Hayashi, in preparation; Hayashi, Nogami, Maeda, & Ikeda, 2004). Examinees can take the tests anytime and anywhere with a computer and the Internet connectivity. Furthermore, they can see the results of the test immediately after taking the test. The WeTEC and CASEC are CATs based on the item response theory (IRT), which adaptive testing typically relies on. It is the standout feature of CAT that an item to be presented to an examinee is selected according to the examinee’s responses to the previous item. Therefore, almost all examinees see different combinations of items according to their responses even though they take the tests at the same time.

Both the WeTEC and CASEC consist of the four sections; vocabulary, conversational schema, listening comprehension and partial dictation. Scores on each of the four sections are reported on a scale from 0 to 250. The scores for the sections are also summed to report a total score which ranges form 0 to 1000. The WeTEC differs from the CASEC mainly in terms of its test length; the WeTEC consists of a total of 100 items, while the CASEC consists of 55 items.

A test question is referred as an “item” in IRT.
Section 1: Section 1 assesses vocabulary knowledge. Examinees complete the presented sentence by selecting the most appropriate word from four alternatives to fill in the blank. The time limit for each item is 60 seconds. 25 items are given in this section.

Section 2: Section 2 assesses knowledge of phrasal expression and usage. Examinees read a given dialogue, look at a picture that describes the situation and then select the most appropriate verbal response from four choices to fill in the blank. The time limit for each item is 90 seconds. 25 items are given in this section.

Section 3: Section 3 assesses listening ability to understand main idea. Examinees listen to passages or dialogues through speakers or headphones and attempt to understand the content. After listening to given passages, examinees are presented with a question and choose the most appropriate answer from four given alternatives. The time limit for each item is 60 seconds. 25 items are given in this section.

Section 4: Section 4 assesses listening ability to understand specific information. This is a dictation task in which examinees listen to sentences being read while observing the written text on their computer screen. Examinees type in the missing words in the blanks within the text. The time limit for each item is 120 seconds. 25 items are presented in this section.

Figure 1: The summary of the WeTEC.

WeTEC is longer than the CASEC in order to keep the measurement accuracy more precise. The properties of the WeTEC are summarized in Figure 1.

3 Methods

In order to examine the relationship between the WeTEC scores and the CEFR levels, we conducted a questionnaire survey.

3.1 Materials

North and Schneider (1998) developed a check list for a learner’s subjective assessment of their ability in English as well as for a tutor’s assessment of his or her student’s proficiency. The list consists of a total of 221 can-do descriptors divided into seven skills and six CEFR levels. The 99 descriptors of the list, which relate to Spoken Production, Spoken Interaction, Strategies and Language Quality, were used in this research (see some examples in the appendix), because those skills are focused in the English Tutorial Lesson. Four types of questionnaires were made of the items according to the class levels that students belong to. The composition of the four questionnaires is shown in Table 3.

The can-do descriptors in the questionnaires were originally written in English. They were translated...
into Japanese and the Japanese-version questionnaires were also prepared, where both Japanese and English descriptors were presented side by side.

3.2 Procedure

The questionnaire survey was conducted during the last fifteen minutes in the last lessons of the English Tutorial Lesson. The CEFR assessment questionnaires were given to both students and tutors. Students were assigned one of the four Japanese versions according to their class levels. They rated themselves for each descriptor on the four-point Likert scale; "4:I can almost do it perfectly", "3:I can do it to a certain degree", "2:I can’t do it very well" and "1:I can hardly do it at all." The English version questionnaires consisting of the same descriptors as those given to the students were assigned to tutors, who are native English speakers. A tutor assessed a student who was randomly chosen from the four students in her/his class. The tutor rated the student using the same Likert scale; however, the tutor might choose “unsure” when she/he could not predict the student’s performance. There were 2619 students who responded to the questionnaires. 982 of them were also assessed by the tutors. The 4-point responses were converted into dichotomous data; 4 and 3 were scored as ‘1’ and 2 and 1 were scored as ‘0.’

Students also took the WeTEC as an achievement test. The time spans between the time they took the test and the time they answered to the questionnaires were less than a month.

3.3 Logistic Regression

In order to predict a probability that a student who took a certain WeTEC score will respond to an item as ‘can,’ the logistic regression analysis was used. Logistic regression is a technique for analyzing and predicting dichotomous variable, say, \( y \), which has only two categorical values such as "can" or "cannot."

Let \( x \) be a predictive variable and \( p(x) \) be a probability of occurrence of the positive outcome, \( y=1 \) or \( y="can", \) given \( x \). The logistic regression models the probability as follows:

\[
p(x) = \frac{1}{1 + \exp(-\alpha - \beta x)}
\]  

where, \( \alpha \) and \( \beta \) are, respectively, the slope and the intercept parameters. Given a set of data, namely, \((y_i, x_i), i=1,2,...,N\), those parameters can be estimated by the method of maximum likelihood.

Let \( \beta^* = \frac{-\beta}{\alpha} \). Then the equation (1) can be rewritten as follows:

\[
p(x) = \frac{1}{1 + \exp(-\alpha(x - \beta^*))}
\]  

The equation (2) is identical to the item response function (IRF) for the two parameter logistic (2PL) model, which is one of the most popular dichotomous models in IRT. In the framework of IRT, the probability that an examinee answers correctly to an item is modeled by IRF with item parameters and examinees.
proficiency parameter. The IRF for the 2PL model is described as follows:

\[ P(\theta) = \frac{1}{1 + \exp\left(-D\alpha_j (\theta - b_j)\right)} , \]

where \( \alpha_j \) and \( b_j \) are discrimination and difficulty parameters for the item \( j \), respectively. The \( \theta \) represents the proficiency parameter for an examinee and the constant \( D \) is typically set to 1.7 (Lord, 1968).

Because the equation (2) and (3) are very similar, the \( \alpha \) and \( \beta^* \) in the equation (2) can be considered as discrimination and difficulty parameters as \( \alpha_j \) and \( b_j \) in the equation (3) are.

For each CEFR item, two sets of \( \alpha \) and \( \beta^* \) for students data and for tutors data were calibrated using the total score of the WeTEC as a predictive variable.

### 3.4 Level Characteristic Curves

In the IRT framework, the test characteristic curve (TCC) is defined as follows:

\[ T(\theta) = \sum_{j=1}^{n} P_j(\theta) , \]

where \( P_j(\theta) \) is an IRF for item \( j \), \( n \) is the number of items in a test. \( T(\theta) \) is the number-right true score that can be interpreted as an expected score of a test given an examinee whose proficiency level is at \( \theta \).

We can also define the same kind of curves for the CEFR levels as 'level characteristic curves (LCCs).'</n>

Let \( l \) be one of the CEFR level and \( s \) is one of the skills. The LCC for level \( l \) of skill \( s \) can be defined as:

\[ T_{ls}(x) = \frac{1}{n_b} \sum_{j=1}^{n_b} p_j(x) , \]

where \( p_j(x) \) is an logistic regression function for item \( j \), \( n_b \) is the number of items in level \( l \) of skill \( s \). Because the number of items in a level of a skill varies among levels or skills, the summation part is divided by the number of items. The \( T(x) \) can be interpreted as a ratio to which a person who got WeTEC score \( x \) will respond as 'can' to the items in the level.

Thus these LCCs relate WeTEC scores to ratios which examinees respond as 'can' to items in each CEFR level. When WeTEC score \( x \) satisfies \( T(x) > .8 \), an examinee who got the score can be seen that she/he would respond as 'can' to more than 80% of the items in the level.

### 4 Results and Discussion

Spearman's rank correlation coefficients between \( \beta^* \)s as item difficulties and CEFR levels was .75 for students data and .93 for tutors data. It can be said that item difficulties assessed by tutors were more consistent to the CEFR levels than difficulties by students. Although both students data and tutors data were analyzed to examine the relationship between the WeTEC scores and CEFR levels, only the results about the relationship between the students' WeTEC scores and the CEFR levels assessed by tutors will be reported in this paper.

Figure 2 shows the LCCs for the four skills obtained from the tutor assessment.

By using the LCCs, we can estimate the proportion of items that a tutor would judge that a student of a certain WeTEC score is able to do. For example, in terms of Spoken Interaction, a student who got 600 points in WeTEC is considered that she/he can perform about 80% of the B1 items, but she/he can perform less than 40% of the C1 items.

It can be said that the LCCs are very useful to see the relationship between the WeTEC scores and the CEFR levels. We can set a cutpoint at which an examinee could be considered to achieve the level. Score \( x \) which satisfies \( T_b(x) = .8 \) can be defined as a threshold for the level \( l \) of skill \( s \) that an examinee could be considered to achieve the level. However, the situation is not so simple. As it is observed in the figure that some LCCs are crossed, there are some pairs that the ranks of cutpoints and the ranks of the CEFR levels are reversed. Further investigation with new data is necessary.

**Acknowledgment**

The authors are sincerely grateful to Dr. Shin-ichi Mayekawa for his valuable advice, and Dr. Tomoko Yamada for her language support.

**References**


Figure 2: The LCCs for the four skills based on the tutor assessment.
Appendix. Sample descriptors of the CEFR assessment questionnaires.

Spoken Interaction
A1 I can introduce somebody and use basic greeting and leave-taking expressions.
A2 I can make simple transactions in shops, post offices or banks.
B1 I can start, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest.
B2 I can initiate, maintain and end discourse naturally with effective turn-taking.
C1 I can keep up with an animated conversation between native speakers.
C2 I can take part effortlessly in all conversations and discussions with native speakers.

Spoken Production
A1 I can give personal information (address, telephone number, nationality, age, family and hobbies).
A2 I can describe myself, my family and other people.
B1 I can give detailed accounts of experiences, describing feelings and reactions.
B2 I can give clear, detailed description on a wide range of subjects related to my fields of interest.
C1 I can give clear, detailed descriptions of complex subjects.
C2 I can present ideas and viewpoints in a very flexible manner in order to give emphasis, to differentiate and to eliminate ambiguity.

Strategies
A1 I can say when I don’t understand.
A2 I can ask for attention.
B1 I can repeat back part of what someone has said to confirm that we understand each other.
B2 I can use standard phrases like “That’s a difficult question to answer” to gain time and keep the turn while formulating what to say.
C1 I can use fluently a variety of appropriate expressions to preface my remarks in order to get the floor, or to gain time and keep the floor while thinking.
C2 I can backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it.

Language Quality
A1 NO DESCRIPTORS
A2 I can make myself understood using memorised phrases and single expressions.

B1 I can keep a conversation going comprehensively, but have to pause to plan and correct what I am saying - especially when I talk freely for longer periods.
B2 I can produce stretches of language with a fairly even tempo; although I can be hesitant as I search for expressions, there are few noticeably long pauses.
C1 I can express myself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.
C2 I can express myself naturally and effortlessly; I only need to pause occasionally in order to select precisely the right words.