Bridging a gap between L2 research and classroom practice (1): English as a Lingua Franca (ELF) in Asia and some assessment based on Common European Framework of Reference for Languages (CEFR)

Michiko Nakano[1], *Eiichiro Tsutsui*[2] *and Yusuke Kondo*[3]

[1] Faculty of Education and Integrated Arts and Sciences, Waseda University
[2] International Center, Hiroshima International University, and
[3]Language Education Center, Ritsumeikan University
nakanom@waseda.jp, ykondo@fc.ritsumei.ac.jp, tsutsui@ic.hirokoku-u.ac.jp

Abstract

Since English has become a lingua franca in the world, English programs need to be based on International standards such as Common European Framework of Reference (CEFR). (1) We report validation experiment by can-do self-check survey in relation to the CEFR levels, in order to find out the cut-off scores and ranges of our computer adaptive placement test called Web-based Test of English Communication (WeTEC), using logistic regression analysis. (2) We discuss Oral self-introduction data among Asian users of English based on the CEFR, using the multi-faceted Rasch Model.

Index Terms: CEFR, logistic regressions, cut-off scores, Rasch model

1. Introduction

Crystal (2003) estimated the number of native speakers of English is 350 million, the number of L2 users of English is 400 million and the number of learners of English as a foreign language is 750 million. Currently, out of 6 billion people in the world, about 1.5 billion people are using or learning English. It amounts to 25 % of the world population. Globalization has changed the world and the way we use English. According to Jenkins (2007, p28), interactions in English between non-native speakers are more common and frequent (80%), while interactions between native speakers of English and non-native speakers are less frequent (20%). English has become well-established lingua franca in the world. With English being the most frequently used lingua franca, most communication happens without the participation of native speakers of English. For this reason, English as a lingua franca (ELF) has been receiving increasing attention in L2 research. The development and use of ELF is probably the most radical and controversial approach to emerge in recent years: see Graddol (2006). It appears to be inevitable that in the use of global English which involves fewer interactions with native speakers, the myth of a standard language becomes more difficult to maintain. Not only in business organizations where English became the corporate language, but also in universities, cyber meetings in English via video-conference systems has become popular and normative: see Nakano and Haraguchi(2009). It is important for us to understand how non-native speakers use English, when talking to other non-native speakers. ELF corpus such as Vienna-Oxford International Corpus of English (VOICE) and AEASOP helps us understand ELF research. Jenkins (2007) has offered ELF core which is the inventory of minimal levels of pronunciation ELF speakers must bear in mind. Seidlehofer (2004) proposed ELF lexico-grammar. Some have viewed their efforts as recommendations towards simplified use of English. This is a naïve view of those who do not know the history of English and the history of morphological simplification as well as syntactic simplification over the years.

At the same time, in order to promote human mobility and to aim at cohesive society where equal opportunities for education and jobs are hoped to achieve, European Council of Education specified the Common European Framework of Reference for languages (CEFR) in 2001 and publicized 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 teaching tasks, and to judge learning achievement. In the globalization age, language proficiency particularly of English as ELF is the necessary commodity for the society to guarantee to future employers the proficiency levels of job candidates and to scale the proficiency levels of prospective employees in terms of pan-European standards. CEFR is widely used in the world as standards to reform curriculum and teaching materials at all levels of formal education. TOEIC, STEP and PhonePass proficiency assessment exams claim to have linkage to CEFR by providing validation experiments. In this paper, we report in Section 2 CEFR-based can-do self-assessment checklist among Japanese university students and compare the result in European Council. In the last section we report our assessment of oral self-introduction among Asian ELF users.

1.1. Can-do Self-check survey and its linkage to CEFR

The English Tutorial Lessons at Waseda University have been offered by the Open Education Center since 1997. In the past six 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 based on the 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. In our previous study in Tsutsui, Kondo and Nakano (2007), Can-do self-check survey data among 2619 students and 982 tutors yielded some evidence that Item characteristic curves based on IRT show the six levels of Tutorial English, following the CEFR six levels. Our goal of the experiment was to show statistically significant correspondence between the outcome of learning CEFR-based textbooks by Waseda students and the tutors' assessment of individual learner performance. We have adopted the questionnaire proposed by North & Schneider (1998). They developed the questionnaire format to elicit a learner's subjective assessment of their ability in English as well as to elicit a tutor's assessment of his or student's ability. We have used 99 items out of 221 originally in North & Schneider (1998). The 99 items relate to Spoken Production, Spoken Interaction, Strategies and Language Quality. 2619 students and 982 tutors took part in the experiment. Table 1 shows the breakdown of participants.

Table 1 Participants

Levels	Participants
Beginner	32 (13)
Basic	417 (153)
Pre-Intermediate	591 (225)
Intermediate	601 (229)
Pre-Advanced	704 (266)
Advanced	274 (96)
Total	2619 (982)

[Number in the brackets indicates the number of participants who are assessed by their tutors.]
We estimated Item Difficulties as well as Item Discriminations according to 2-parameter logistic model:

$$p_j(\theta) = \frac{1}{1+\exp\left(-Da_j(\theta - b_j)\right)} \quad (1)$$

Where a_j stands for Item Discrimination and b_j, Item Difficulty..

Table 2 shows the correlations of Item Difficulties between those by students and by tutors.

Table 2 Correlation of item difficulties

	r
Spoken Interaction	.849
Spoken Production	.969
Language Strategies	.899
Language Quality	.831

Fig 1 represents students' assessment of Item Difficulties for Spoken Interactions; Fig 2, for Spoken Productions; Fig 3, for Language Quality. Apart from B1 in Strategies and C1 in Language Quality, Item Characteristic Curves are not crossed. This indicates that teaching materials correspond to the levels of descriptors in CEFR on the whole.

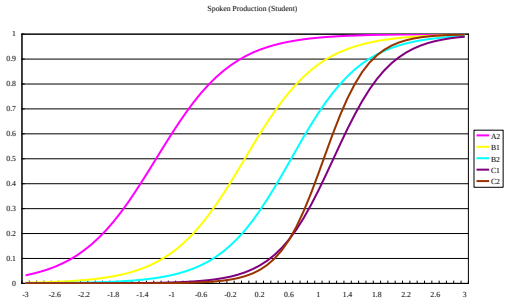


Figure 1. Item Difficulties for Spoken Interactions (students)

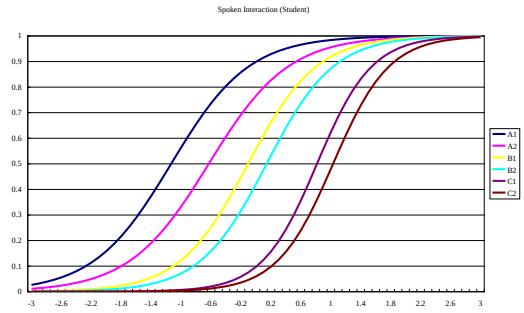


Figure 2. Item Difficulties for Spoken Productions (students)

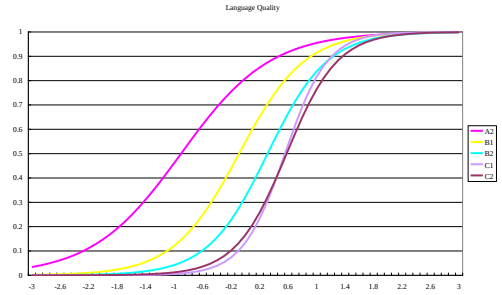


Figure 3. Item Difficulties for Language Quality (students)

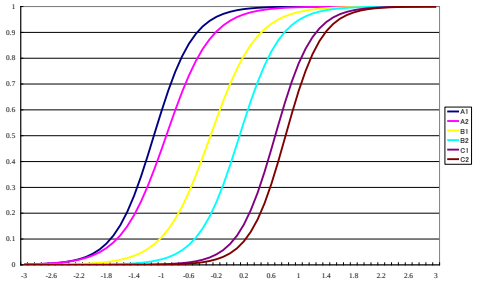


Figure 4. Item Difficulties (Tutors)

Fig 4 shows Tutors' assessment of Item Difficulties. Although the Item Discriminations are sharper and more distinct than those by the students, none of the curves are crossed. This indicates that tutors regards teaching materials as following the descriptor levels specified by CEFR.

1.2. Relating WeTEC to CEFR

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 is hoped that correspondences between the WeTEC scores and the CEFR can be obtained. In this research, four formats of the CEFR can-do self-check questionnaires were given to both students and tutors. This section describes our attempt to link the WeTEC scores and

the CEFR levels, analyzing the data by logistic regressions.

1.2.1. 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 outcomes where a variable has only two values such as "can" or "cannot."

Let p be the probability of occurrence of an event, $p/(1-p)$ is called as 'odds' and its logarithm $\log(p/(1-p))$ is referred as 'logit.' Logistic regression analysis provides prediction of the probability of occurrence of an event by modeling the logit as a linear function of a predictive variable.

Let x be a predictive variable and $p(x)$ be a probability of occurrence of an event given x. The model to be fitted is described as follows:

$$\log \frac{p(x)}{1-p(x)} = \beta + \alpha\,, \qquad (2)$$

where, α and β are regression coefficients.

The predictive function of $p(x)$ can be derived by modifying the equation (2):

$$p(x) = \frac{1}{1+\exp(-(\beta+\alpha x))} \qquad (3)$$

Let $\beta^* = \frac{-\beta}{\alpha}$. Then the equation (2) can be rewritten as follows:

$$p(x) = \frac{1}{1+\exp(-\alpha(x-\beta^*))}. \qquad (4)$$

The equation (3) looks very similar 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 in (1).

For each CEFR item, two sets of α and β^* for students data and for tutors data were calibrated using the total score of the WeTEC as a predictive variable.

1.2.2 Level characteristic curves

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

$$T(\theta) = \sum_{j=1}^{n} p_j(\theta), \qquad (5)$$

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 θ.

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

$$T_{ls}(x) = \frac{1}{n_{ls}} \sum_{j=1}^{n_{ls}} p_j(x), \qquad (6)$$

where $p_j(x)$ is an logistic regression function for item j, n_{ls} 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.

1.3. Result and discussion

The correlation coefficients between β^*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 are reported during the presentation.

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 cut-off point at which an examinee could be considered to achieve the level. Score x which satisfies $T_{ls}(x) = .8$ 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.

2. Oral Self introduction and its assessment by CEFR

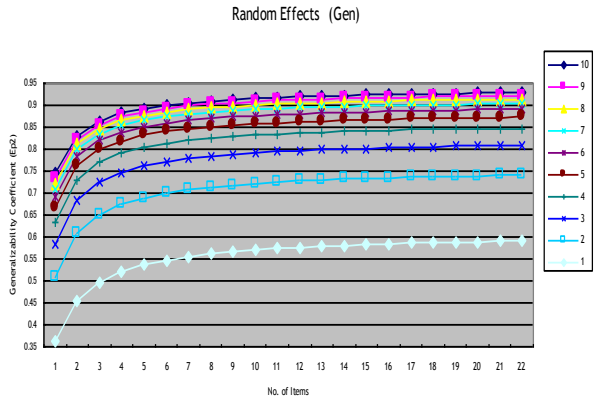
The purpose of our experiment is to assess oral self-introduction in English in reference to CEFR, in order to validate the practicability of CEFR for the assessment of Asian English learners. The present assessment of oral English among Asian users of English was jointly analyzed by Michiko Nakano, Eichiro Tsutsui and Yusuke Kondo in 2006. The oral speech data was here reanalyzed by multi-faced Rasch model.

10 human raters rated 73 Asian learners of English, based on CEFR, with respect to the three major categories of Intelligibility, Comprehensibility and Interpretability. We also evaluated their speaking abilities by 24 subordinate items: see details Nakano and Tsutsui (2010). At the same time, we assessed JACET 8000 word levels, C-units, error-free C-units, type-token ratio, frequency of filled pauses, silent pauses, articulation rates, wpm, average length filled pauses, average length of silent pauses, Rate of filled pauses (Rate of filled

pauses per length of time of utterance) and Rate of silent pauses (Rate of silent pauses per length of time of utterance).

2.1 Result and discussion

By using FACETS, we can estimate (1) examinee's ability, (2) rater's severity and (3) item difficulty. Our initial analysis was defective in that, although Outfit mean square shows that the data fits the Rasch Model, since every value is within the range of 0.9 to 1.1 and below 2.0., step difficulty index did not satisfy the criteria of step difference (more than 1.4 logits and less than 5.0 logits), suggesting there is some difference in rater severity among our 10 raters. Another drawback of our initial model is that there are several misfit items that should be excluded from our analysis. "Absence of tension" and "Absence of foreign accentedness" overly exceeded the criteria of Infit MS Mean ±2.0SD: see details in Kondo-Brown, 2003. Examinee's abilities should be re-estimated by using the newly selected items. Furthermore, Inter-Rater agreement opportunities are 78840, and the exact agreements are 22425 out of 78840 (28.4%). This suggests some difference in rater severity and consistency. For this reason, we analyzed the data again by FACETS. G-coefficients (Brennan, 2000) were calculated to see how many items and how many raters if they are reliable by severity index we need for the further analysis.



This statistics shows that we do not need 22 items and ten raters. We only need four or five items and four or five raters. In terms of rater severity, Table 3 indicates that four raters are most reliable.

Table 3 Rater measurement report

	Logit	InfitMS
Judge1	**0.51**	1.1
Judge2	-0.8	0.9
~~Judge3~~	~~-0.3~~	~~1.3~~
~~Judge4~~	~~-0.3~~	~~1.4~~
Judge5	-0.4	1.2
~~Judge6~~	~~**-1.5**~~	~~1.6~~
~~Judge7~~	~~0.04~~	~~0.4~~
Judge8	-0.6	1
~~Judge9~~	~~0.18~~	~~0.5~~
~~Judge10~~	~~-0.2~~	~~0.6~~

Therefore, by eliminating misfit raters and items, the estimations were conducted until the criterion of infitMS Mean ±2.0SD is reached. In the process of estimation, five items were found to be preferably excluded from the analysis.

The result then shows that step difference is also within the range of satisfactory criteria (more than 1.4 logits). In our present data, reliability coefficients of our participants, raters and items became 0.99, 1.00 and 0.96. Inter-rater exact agreements are 2979 out of 8322 opportunities, which are slightly increased to 35.8% from the initial 28.4%.

Table 4 Four rater's measure

	R^2	F	B
WPM	.78	56.7**	-.99**
Ratio of Filled Pause			-.27**
TTR			-.26**

**p < .01, two-tailed. N = 30. df = 2, 27

Overall results show that

- 6-point scale ratings based on CEFR are applicable for assessing speech in English among Asian users of English.
- 6 proficiency levels by our rating procedures are consistent with the results of our objective measurements. In particular, G-statistics shows that we only need four raters and four or five items.
- Step-wise multiple regression analysis shows that abilities estimated by FACETS are reliably predictable by three objective measurements of Word per Minute, Ratio of Filled Pauses and Type-token Ratio. These three items explained 78% variance of our data (Table 4).

The rescaled item difficulties are all within the range of North and Schneider's (1998) estimates of the cut-off logit scores based on the FACETS analysis among 2865 learners.

3. References

[1] Crystal, D. (2003). *English as a Global Language*. 2nd ed. CUP.

[2] Jenkins, J. (2000). *The Phonology of English as an International Language*. OUP.

[3] Graddol, D. (2006). *English Next*. The British Council.

[4] Jenkins, J. (2000). *The Phonology of English as an International Language*. OUP.

[5] Nakano, M. and Haraguchi, Y. Cyber course on World Englishes and ELF: some tentative evidence. *Proceedings of the 14th Conference of Pan-Pacific Association of Applied Linguistics*, 385-392.

[6] Seidlehofer, B. (2004). Research perspective on teaching English as a lingua franca. *Annual Review of Applied Linguistics* 24. 209-239.

[7] Tsutsui, E., Nakano, M., & Kondo, Y. (2007). Nihonjin eigo gakushuusha no jissen-teki hatsuwa nouryoku ni kansuru hyouka kijun no kentou: Common European Framework of References wo kiban to shite. [An investigation on an assessment criterion for practical speech proficiency of Japanese English learners: A research based on the Common European Framework of References.], *Proceedings of the 5th Annual Conference of The Japan Association for Research on Testing*, 88-91.

[8] North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 15(2), 217-263

[9] Nakano, M. and Tsutsui, E. Three models of World Englishes and our personal perspective. *JACET-ICT Practice and Research 2009*. pp195-210.

[10] Brown, L. and McNamara, T. F. (1998). Using G-theory and Many-facted Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 20. 15(2), 1-25.

[11] Brennan, R.L. (2001). *Generalizability Theory*. Springer.