Prediction of General ESL Proficiency Considering Learners' Dictation Performance

Katsunori Kotani[1*], and *Takehiko* Yoshimi[2]

[1]Kansai Gaidai University, College of International Professional Development, Osaka, Japan
[2]Ryukoku University, Faculty of Advanced Science and Technology, Shiga, Japan

Abstract. This study analyzes the extent to which dictation performance and linguistic features (linguistic difficulty of sentences during dictation) can predict general proficiency in English as a second language (ESL) learners. To this end, this study constructed a multiple linear and a non-linear regression models that predict general ESL proficiency (in which independent variables were the dictation performance scores and the linguistic features of sentences) and verified the correlation between the predicted and observed general ESL proficiencies. The results showed that general ESL proficiency could be predicted by dictation performance and linguistic features. Furthermore, the results indicated significant effects on dictation accuracy, sentence length, and mean word length.

1 Introduction

During English as a second language (ESL) education, teachers evaluate learners' learning outcomes over a semester by using proficiency tests at the beginning and end of the semester. However, using proficiency tests has its own problems: test fees are expensive, and the test administration takes a long time, for example, the Test of English for International Communication (TOEIC) is two hours long. The cost has been decreasing because of the availability of computer-based tests, which replace paper-based tests. However, it is still difficult to assign computer-based proficiency tests repeatedly in a semester, because the cost is not easily affordable.

Computer-based proficiency tests, for example, Duolingo and Versant, measure general ESL proficiency. Among the different test items, these tests use a dictation test. In addition, a dictation test was used to measure the general proficiency of ESL [1, 2]. Dictation test results correlate with general ESL proficiency because learners carry out dictation by using their phonetic/phonological ability for recognizing a series of phonemes, and morphological/syntactic/semantic ability for constructing words, phrases, and sentences.

Previous research [2, 3, 4, 5, 6] presumed that dictation performance is a good indicator of general ESL proficiency, and compared learners' dictation test scores with general ESL proficiency scores such as the TOEIC and the Test of English as a Foreign Language (TOEFL). The results showed a strong correlation between the dictation test scores and

* Corresponding author: kkotani@kansaigaidai.ac.jp

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

general ESL test scores. Hence, previous research has succeeded in demonstrating that dictation performance indicates observed test scores for general ESL proficiency.

This study addresses three major questions that previous research has not resolved. First, previous research demonstrated theoretical evidence for the dictation-based model for ESL proficiency by using a simple correlation analysis in a close test. Here comes a question whether the dictation-model is adequate when ESL teachers use this model in their classes. That is, the dictation-model should also be empirically verified. Therefore, this study examines the adequacy of the dictation-based model for ESL proficiency in an open test, using non-linear regression analysis.

Second, previous research determined learners' dictation performance based on the accuracy of dictation, that is, how correctly learners wrote down what they heard. The high dictation accuracy includes different proficiency levels. One is the "true" high proficiency level, and the other is the pseudo-high proficiency level. Learners at the high proficiency level complete dictation correctly and easily, and it is not a strenuous task for them. Conversely, learners at the pseudo-high proficiency level complete dictation correctly, but they find the dictation task difficult, and it is strenuous for them. This study evaluates the dictation performance not only based on accuracy but also on the students' ease for further distinguishing the accuracy-based proficiency.

Third, previous research failed to include sentence difficulty such as the sentence/word length in developing a dictation-based model, although sentence difficulty would affect the measurement of ESL proficiency. This study distinguishes high proficiency from the viewpoint of sentence difficulty. Here, learners at the high proficiency level complete dictation correctly even for difficult sentences. However, those at the pseudo-high proficiency level succeed in completing dictation for easy sentences, but fail to do so when given difficult sentences. As general ESL proficiency depends on the linguistic difficulty of sentences [7], this study evaluates it considering the linguistic features of sentences.

To fill these gaps in the literature, this study aims to analyze the extent to which dictation performance and linguistic features can predict general ESL proficiency from the assessment of dictation performance using two evaluation criteria (namely, ease of dictation and dictation accuracy). More specifically, this study addresses the following research questions:

- RQ1: To what extent can general ESL proficiency be predicted by composite dictation performance and linguistic features?
- RQ2: What are the effective predictors of general ESL proficiency?

To this end, this study constructed a multiple linear and a non-linear regression models that predict general ESL proficiency and verified the correlation between the predicted and observed general ESL proficiencies.

2 Methods

2.1 Participants

The participants of this study were 50 English learners. This number was determined to mimic a large English class that includes learners at different proficiency levels. This is because this study placed more emphasis on the practical application of model building than on the theoretical perspective. In addition, the participants were not randomly chosen. Those

who satisfied the following conditions participated in the experiment: their first language was Japanese; they were students of universities in the area where this study was carried out (28 men and 22 women; mean age, 20.8 years; standard deviation (*SD*), 1.3). The participants were paid a fee for participation.

2.2 Dictation data collection

Dictation data were compiled following Kotani and Yoshimi [8]. Data instances to determine general ESL proficiency comprised sentences transcribed by a learner, two types of dictation performance scores, five types of linguistic features extracted from reference sentences from textual material, and the learners' English test scores. The dictation data included 750 instances gathered from 50 learners' attempts to complete a textual material consisting of 15 sentences.

The dictation task proceeded as follows: First, the 50 learners listened to sentences read aloud by a voice actor (woman, 35 years old) who was a native speaker of American English, and transcribed them sentence-by-sentence. Subsequently, the learners subjectively judged their ease of dictation (explained in section 2.5). Three instructions were given to the learners: 1) each sentence could be read twice, if necessary; 2) each task should be completed at a natural speed for the learner; and 3) the learners were forbidden to return and revise a sentence after moving on to the next one, regardless of available time.

2.3 Text material

Two types of texts were selected from those distributed by the International Phonetic Association [9] and Deterding [10]. As these texts include basic English sounds, an analysis of the learners' dictation of these texts would reveal what types of English sounds influenced their listening. These texts featured two of Aesop's Fables: The North Wind and the Sun (Text I) and The Boy Who Cried Wolf (Text II). Texts I and II contained five and ten sentences, respectively. It is noteworthy that Text I failed to encompass certain sounds, such as initial and medial /z/ and syllable initial /θ/. Accordingly, Text II covered these missing sounds.

2.4 General English proficiency

Learners' general English proficiency (GEP) was determined using their TOEIC Listening & Reading test scores, obtained in the current or previous year. The TOEIC Listening & Reading test was chosen, because the test scores had strongly correlated with GEP test results, that is, the Language Proficiency Interview developed at the Foreign Service Institute of U.S. Department of State [11], and this test has no dictation section.

2.5 Dictation performance

The criteria for evaluating dictation performance comprised two indexes: learners' subjective judgment of their ease with dictation (EASE) and dictation accuracy (ACC).

EASE was scored using a five-point Likert scale for the learners' subjective judgment (1: easy; 2: somewhat easy; 3: average; 4: somewhat difficult; and 5: difficult). A lower EASE indicated that the learners judged the dictation to be easier.

ACC was calculated by dividing the Levenshtein edit distance between a given reference and a transcribed sentences with the number of characters in a longer sentence than the other. The Levenshtein edit distance reflects the differences between the two sentences due to the substitution, deletion, or insertion of characters. A lower ACC denoted that the learners completed the dictation more accurately.

2.6 Linguistic features: Linguistic difficulty of sentences

In this study, linguistic features included sentence length [12], mean word length [12], the number of multiple-syllable words [13, 14], word difficulty [15], and speech rate [16].

The sentence length was defined as the number of words in a sentence.

Mean word length was derived by dividing the number of syllables by the number of words in the sentence. The number of syllables in a given word was counted using the following steps: Count the vowels in the word, subtract any silent vowels, and subtract one vowel from every diphthong [17].

The number of multiple-syllable words in a sentence, Fang's listening score [13], was derived by formula (1), where N denotes the number of words in the sentence, and S_i denotes the number of syllables in the i-th word. This subtraction derivation ignored single-syllable words.

$$\sum_{i=1}^{N}(S_i - 1) \quad (1)$$

Kiyokawa's word difficulty was defined as the rate of words not listed in his list of basic spoken words in relation to the total number of words in the sentence.

The speech rate was defined as the number of words read aloud by the native speaker in one minute.

These linguistic features were simple features that would effectively measure the difficulty/ease of the listening comprehension of a sentence. These features were automatically derived from the sentences in the text material.

2.7 Prediction of general ESL proficiency with dictation performance and linguistic difficulty of sentences

Prediction models were developed using a non-linear regression model and a multiple linear regression model. The dependent variable was GEP, and the independent variables were the dictation performance scores (EASE, ACC) and the linguistic features described in Sections 2.5 and 2.6.

To answer the RQ1, non-linear regression using support vector machines [18] was conducted to determine the extent to which dictation performance and linguistic features predicted general ESL proficiency. Support vector regression was carried out using the function "svm()" as defined in the "e1071" package of the software environment R [19]. The radial basis function was set as a type of kernel function, and the other parameter settings of "svm()" were set at their default. See pp.49–53 of [19] for details of the default settings.

The overall effect of the dictation performance scores was evaluated in a leave-one-out cross-validation test, taking one instance as test data and n-1 instances as training data (n = 750). A leave-one-out cross-validation test is preferable for small datasets because it creates

the largest possible test set for a fixed training dataset [20]. Correlation analysis was carried out between the predicted and observed GEPs.

Correlations of GEP with EASE/ACC were analyzed to investigate how well the prediction models predict GEP. The correlation analyses used learners' mean EASE/ACC scores, that is, a learner's sum score of EASE/ACC divided by the number of sentences (n = 15).

To answer the RQ2, a multiple linear regression was conducted to determine the effective predictors of general ESL proficiency. Although higher performance has been verified with support vector machines, multiple linear regression was employed to evaluate the effects of the dictation performance scores and linguistic features. Before the multiple linear regression analysis, the independent variables were examined in terms of the presence of multicollinearity by calculating the variance inflation factor (*VIF*) [21].

The effects of the dictation performance scores and linguistic features were evaluated using standardized partial regression coefficients. It was found that the effects increase with the absolute value of the coefficients.

3 Results

3.1 Descriptive statistics

Figure 1 shows the distribution of GEP. GEP followed a normal distribution according to the Kolmogorov-Smirnov test (K = 0.82, p = 0.25). Table 1 shows the descriptive statistics of GEP, the dictation performance scores, and the linguistic difficulty of the sentences.

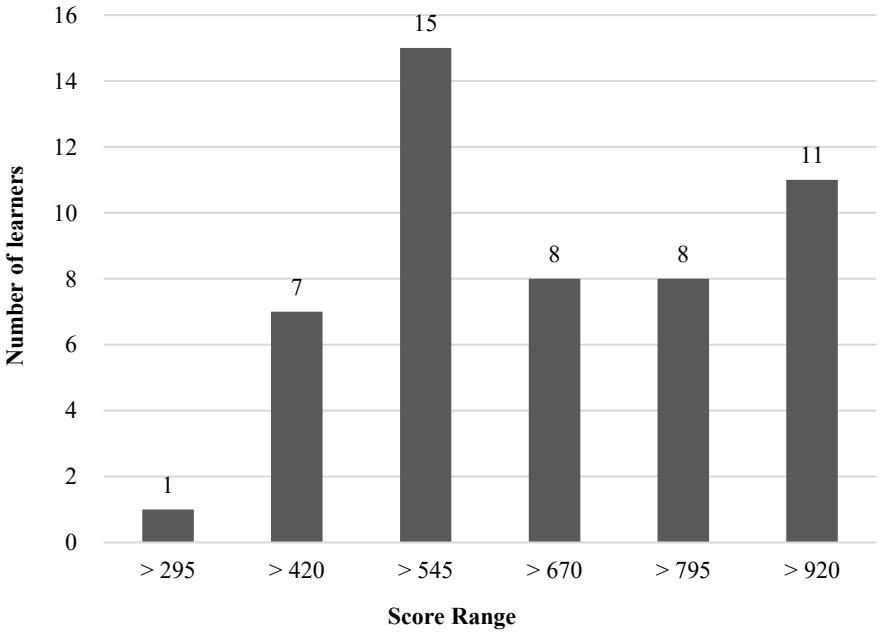

Fig. 1. GEP distribution.

Table 1. Descriptive statistics of GEP and dictation data.

Feature type	*n*	Mean	*SD*
GEP	50	607.7	186.2
EASE	750	4.2	0.8
ACC	750	0.4	0.2
Sentence length	15	21.9	7.6
Mean word length	15	1.3	0.1
Number of multiple-syllable words	15	5.9	2.8
Word difficulty	15	0.3	0.1
Speech rate	15	178.4	17.4

3.2 Result of prediction of general ESL proficiency

Pearson's correlation analysis showed a correlation of large effect size ($r = 0.75$) between the predicted and observed GEPs in the cross-validation test. Figure 2 shows a scatter plot. The correlation analysis showed correlations of large effect size between the observed GEPs and EASEs ($r = -0.51$) and between the observed GEPs and ACCs ($r = -0.84$).

The effect sizes of the predicted GEPs were compared with those of the EASEs and ACCs by comparing the absolute correlation coefficients. The difference of the absolute correlation coefficients showed statistically significant differences between 0.75 and 0.51 ($diff = 0.24$, $z = -2.73$, $p < 0.05$), and between 0.75 and 0.84 ($diff = 0.09$, $z = 1.65$, $p = 0.05$).

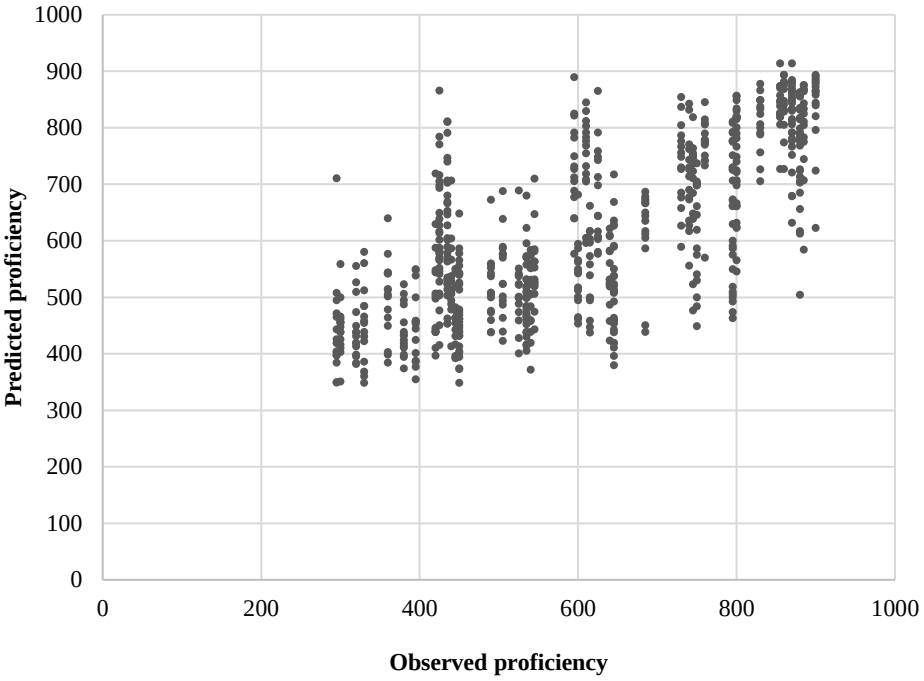

Fig. 2. Scatter plot of the predicted and observed GEP.

Table 2. Standardized partial regression coefficients of the independent variables.

Independent variable	Standardized partial regression coefficient
EASE	0.00
ACC	−0.79*
Sentence length	0.32*
Mean word length	0.12*
Word difficulty	0.06
Speech rate	−0.03

*: $p < 0.05$

Multicollinearity ($VIF > 10$) was observed in the number of multiple-syllable words ($VIF = 12.3$); hence, it was excluded from the independent variables. The multiple linear regression yielded a significant regression equation (F (6, 743) = 125.80, $p < 0.05$), with an adjusted squared correlation coefficient R^2 of 0.50. Table 2 shows the standardized partial regression coefficients. A statistically significant effect was observed for ACC, sentence length, mean word length, and number of multiple-syllable words ($p < 0.05$). The degree ranged in this order.

4 Discussion

The result of the non-linear regression showed a correlation of large effect size between dictation performance and GEP (Figure 2). In addition, this study confirmed a similar effect in an open test, while previous studies examined the correlation in closed tests.

The absolute correlation coefficient of the composite evaluation criteria was significantly higher than that of the single evaluation criterion, EASE, and lower than that of another single evaluation criterion, ACC. The smaller effect size of EASE occurs because of a potential problem due to subjective judgment. This is because a larger effect size was observed in the case of ACC, which is based on objective judgment. For example, learners' judgment of the ease of dictation may depend on their individuality, language aptitudes, and learning backgrounds.

While the EASE-based GEP dictation model showed the least prediction performance, EASE demonstrated no contribution to the prediction. This lack of contribution can be explained as follows. First, it concerns the reliability of subjective judgment on a Likert scale, as mentioned above. Next, it concerns the natural speech rate (178.4 WPM). The natural speech rate was faster than the speech rate that learners listened to for English texts. Thus, the learners might have inappropriately judged the EASE. Finally, it concerns learners' unfamiliarity with the dictation task. Most learners might have had little chance to complete dictation, and the dictation task might have been strenuous for them. This was seen in the inclined distribution of the mean score for EASE, 4.2, which was inclined to the difficult on a five-point Likert scale where scale 5 was the most difficult. Contrarily, such inclination was not observed in the mean score for the ACC, 0.4.

Among the linguistic features, a significant effect was observed for sentence length and mean word length (Table 2). Therefore, the GEP dictation model should be developed taking into consideration of linguistic properties. Although the failure of word difficulty and speech rate should be explored in future studies, the results suggest the necessity of phonetic/phonological features in predicting GEP.

5 Conclusions

This study predicted general ESL proficiency by a multiple linear regression analysis and a non-linear regression analysis using dictation performance scores and linguistic features; it addressed research questions regarding the overall effect of learners' dictation performance and linguistic features, as well as effective predictors of general ESL proficiency. Dictation performance was assessed using two evaluation criteria: EASE and ACC. A statistically significant effect was found for ACC, but not for EASE, word difficulty, or speech rate.

Future studies should examine the prediction accuracy while considering linguistic difficulty and GEP levels. The linguistic features should comprise phonetic/phonological features for listening difficulty, syntactic difficulty for sentence difficulty, and lexical features for spelling difficulty. In addition, future studies should evaluate learners' performance using different methods such as reading-aloud and shadowing, because the reading-aloud is another popular method used in computer-based proficiency tests, and shadowing is a well-recognized, popular method of learning ESL.

The authors acknowledge the reviewers and the participants in ETLTC2021 for their constructive comments on the manuscript and the presentation. This work was supported by JSPS KAKENHI (Grant Numbers 22300299, 15H02940, and 17K18679).

References

1. J. W. Oller Jr. *Dictation as a device for testing foreign language proficiency*. English Language Teaching, **15**, 254-259 (1971)
2. A. Wong and P. Leeming. *Using dictation to measure language proficiency*. Language Education in Asia, **5**, 1, 160-169 (2014)
3. A. Yazdinejad and M. Zeraatpishe. *Investigating the validity of partial dictation as a test of overall language proficiency*. International Journal of Language Testing, **9**, 2, 44-55 (2019)
4. P. Leeming and A. Wong. *Using dictation to measure language proficiency: A Rasch analysis*. Language Testing and Assessment, 5, 2, 1-25 (2016)
5. P. Irvine, P. Atai, and J. W. Oller Jr. *Cloze, Dictation, and the Test of English as a Foreign Language*. Language Learning, 24, 2, 245-252 (1974)
6. S. Kazazoğlu. *Dictation as a language learning tool*. Procedia-Social and Behavioral Sciences, **70**, 1338-1346 (2013)
7. M. Kanzaki. *Minimal English Test: Item analysis and comparison with TOEIC scores*. SHIKEN, 19, 2, 12-23 (2015)
8. K. Kotani and T. Yoshimi. *Effectiveness of Linguistic and Learner Features for Listenability Measurement Using a Decision Tree Classifier*. The Journal of Information and Systems in Education, **16**, 1, 7-11 (2017)
9. International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press (1999)
10. D. Deterding. *The North Wind versus a Wolf: Short texts for the description and measurement of English pronunciation*. Journal of the International Phonetic Association, **36**, 2, 187-196 (2006)
11. Chauncey Group International. *TOEIC Technical Manual*. Princeton, NJ: Chauncey Group International (1998)
12. J. S. Chall and H. E. Dial. *Predicting listener understanding and interest in newscasts*. Educational Research Bulletin, **27**, 6, 141-153+168 (1948)
13. I. E. Fang. *The Easy Listening Formula*. Journal of Broadcasting, **11**, 1, 63-68 (1966)
14. R. Remus. *Improving Sentence-level Subjectivity Classification through Readability Measurement*. Proceedings of the 18th Nordic Conference of Computational Linguistics, 168-174 (2011)
15. H. Kiyokawa. *A formula for predicting listenability: The listenability of English language materials 2*. Wayo Women's University Language and Literature, **24**, 57-74 (1990)
16. J. Messerklinger. *Listenability*. Center for English Language Education Journal, **14**, 56-70 (2006)
17. A. Stenton. *The role of the syllable in foreign language learning: Improving oral production through dual-coded sound-synchronised typographic annotations*. Language Learning in Higher Education: Journal of the European Confederation of Language Centres in Higher Education (CercleS), **2**, 1, 145-161 (2013)

18. V. N. Vapnik. *Statistical Learning Theory*. New York: John Wiley & Sons (1998)
19. D. Meyer, E. Dimitriadou, K. Hornik et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, https://cran.r-project.org/web/packages/e1071 (2018)
20. R. B. Rao, G. Fung, and R. Rosales. *On the dangers of cross-validation: An experimental evaluation*. Proceedings of the 2008 SIAM International Conference on Data Mining, 588-596 (2008)
21. J. Neter, M. Kutner, W. Wasserman, and C. Nachtsheim. *Applied Linear Statistical Models (4th ed.)*. New York: McGraw Hill (1996)