A Comparison of the Validity of Measurement Methods for the General English Proficiency through Dictation and Read-Aloud Performances

Anonymous ACL submission

Abstract

This paper compares three measurement methods for the general proficiency of learners of English as a second language (GEP). If students' GEP can be measured on course materials frequently, for instance, at the beginning and end of a semester, English teachers can confirm students' levels of learning achievement. So far, English teachers have two options for GEP measurement: calculating scores for read-aloud or those of dictation performance. This study expands an option to measure GEP using both dictation and read-aloud performances. When comparing the three types of measurement methods, the experimental results suggest that GEP should be measured by calculating dictation and read-aloud performances.

1 Introduction

Evaluating general English proficiency (GEP), which includes listening, speaking, reading, and writing, is an essential task for teachers of English as a second language. One of the goals of GEP evaluation is to investigate a learner's learning outcome by comparing their GEP at the beginning and end of a semester. GEP evaluation can be conducted using English tests such as the Test of English as a Foreign Language and Test of English for International Communication (TOEIC), because the use of these tests reduces teachers' time and effort in test preparation.

However, when these tests are used as a classroom-based assessment, that is, an evaluation of learners' performance by teachers, there are three limitations. First, the test content is irrelevant to learners' classes. Second, the test duration requires a couple hours more than the class period. Finally, test fees are expensive, as learners must take a test at least twice a semester.

A solution to these limitations is to introduce computer-assisted language testing (Noijons, 1994; Suvorov, 2013). Here, GEP is measured by calculating scores for a learner's read-aloud and dictation performance.

Previous research has been classified into two categories. One examined the correlation of GEP with dictation or read-aloud performances (Irvine, Atai, and Oller, Jr. 1974; Iino, Yabuta, and Thomas 2011; Kanzaki 2015; Leeming and Wong 2016). The other developed a measurement method for GEP based on dictation or read-aloud performances (Kotani and Yoshimi 2021a; Kotani and Yoshimi 2021b). Kotani and Yoshimi (2021a) and Kotani and Yoshimi (2021b) measured GEP using dictation performance and read-aloud performance, respectively.

This study expands a teacher's option by providing the third GEP measurement method of using both dictation and read-aloud performances. Previous research (Kotani and Yoshimi 2021a/b) did not examine the extent to which the measurement performance can be improvement by measuring GEP based on the third method.

The goal of this study is to determine an effective GEP-measurement method by comparing different patterns of sub-proficiencies. Hence, the research question is as follows:

- Which is the highest GEP-measurement performance among a dictation-based method, a read-aloud method, and a dictation and read-aloud-based method?

These three methods are compared not only regarding the measurement accuracy but also the ease of measurement, specifically, the cost of developing and administering a method.

The contributions of the present study are to (1) investigate an effective GEP-measurement method as a classroom-based assessment alternative to GEP tests, (2) examine the validity of GEP-measurement methods, namely, a dictation-based method, a read-aloud method, and both dictation and read-aloud methods, and (3) verify the

robustness of a GEP-measurement method against the English-language-class size of training data.

2 Collection of Dictation and Read-Aloud Data

Data instances comprised sentences transcribed by a learner, two types of dictation performance scores, speech sound pronounced by a learner, three types of read-aloud performance scores, five types of linguistic features extracted from reference sentences in a text material, and the learners' English test scores. The dictation and read-aloud data included 750 instances (50 learners' tasks for 15 sentences).

2.1 Participants

The study participants were 50 English learners. This number was determined to mimic a large English class that included learners at different proficiency levels. The use of class-size training data reveals the possibility of teachers developing a GEP-measurement method using training data compiled in the class.

Participants were not randomly chosen. Those who satisfied the following conditions participated in the experiment: their first language was Japanese, and they were students at universities in the area where this study was conducted (28 men and 22 women; mean age, 20.8 years; standard deviation [*SD*], 1.3). The participants were paid a fee for the experiment.

2.2 Data Collection Procedures

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 (see Section 3.1).

The read-aloud task was performed as follows: First, the learners listened to a reference speech sound by the native speaker. Subsequently, they read a sentence aloud and subjectively judged the ease of reading aloud (see Section 3.2). Their read-aloud durations were recorded to calculate the speech rates.

Learners received three points of instruction: 1) Each sentence could be listened to or read twice, if necessary; 2) Each task should be completed at a speed natural for the learner; 3) It was forbidden to read fast or slowly or to return and revise a sentence after moving on to the next sentence.

2.3 Text Material

Two types of texts were selected from those distributed by the International Phonetic Association (1999) and Deterding (2006). As these texts include basic English sounds, an analysis of the learners' dictation and read-aloud performance of these texts would reveal what types of English sounds influenced their listening and pronunciation.

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. Text I failed to encompass certain sounds, such as the initial and medial /z/ and syllable initial /θ/. However, Text II included these missing sounds.

2.4 General English Proficiency

GEP was determined using participants' TOEIC listening and reading test scores obtained in the current or previous year. The reasons for this choice were as follows: The test scores were strongly correlated with the GEP test results, specifically, the Language Proficiency Interview developed at the Foreign Service Institute of the U.S. Department of State (Educational Testing Service 1998), and that this test has no dictation or read-aloud sections.

3 Features for Regression

3.1 Dictation Performance

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

EASE-D 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).

ACC-D was calculated by dividing the Levenshtein edit distance between a given reference and a transcribed sentence by the number of characters in a longer sentence than the other. The Levenshtein edit distance reflects the differences between the two sentences because of the substitution, deletion, or character insertion.

3.2 Read-Aloud Performance

The criteria for evaluating read-aloud performance comprised three indices: learners' subjective judgment of the ease of reading aloud (EASE-R), read-aloud accuracy (ACC-R), and speech rate in words per minute (RATE-R).

The EASE-R was determined by the learner's subjective judgment on a five-point Likert scale.

ACC-R was calculated by dividing the number of words correctly read aloud by the number of words in the corresponding sentence. A native English speaker evaluated learners' reading aloud word-by-word, but not phoneme-by-phoneme, using a binary decision (correct or incorrect pronunciation). The evaluator was trained to replicate interviews and meetings but was not familiar with the English spoken by learners. Before the assessment, the evaluator read the reference texts.

RATE-R was calculated by dividing the number of words by the duration of reading aloud.

3.3 Linguistic Features

In this study, linguistic features included sentence length, mean word length, number of multiple-syllable words, and word difficulty.

Sentence length (Chall and Dial 1948) was defined as the number of words in a sentence.

The mean word length (Chall and Dial 1948) was derived by dividing the number of syllables by the number of words in the sentence. The number of syllables in a word (Stenton 2013) was counted using the following steps: count the vowels in the word, subtract any silent vowels, and subtract one vowel from every diphthong.

The number of multiple-syllable words in a sentence (Fang 1966) was derived using the formula $\sum_{i=1}^{N}(S_i - 1)$, 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 ignores the single-syllable words.

Word difficulty (Kiyokawa 1990) was defined as the rate of words not listed in Kiyokawa's basic vocabulary list relating 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.

4 Measurement of GEP with Dictation and/or Read-Aloud Performances

The measurement methods were developed using support vector regression, with GEP as the dependent variable. The independent variables were dictation performance scores, read-aloud performance scores, and linguistic features.

Support vector regression was conducted using the function "svm()" defined in the "e1071" package of the software environment R (Meyer 2021). The radial basis function was set as a type of kernel function, and the other parameters of "svm()" were set as default.

The measurement methods were evaluated using a leave-one-out cross-validation test. The training/test data consisted of 750 instances.

A correlation analysis was performed between the measured and observed GEPs. The significance threshold was adjusted for multiple testing based on the false discovery rate (FDR) (Benjamini and Hochberg 1995). A statistically significant correlation was examined to answer the research question.

To address the research question, three types of measurement methods were developed: dictation performance scores, read-aloud performance scores, and dictation and read-aloud performance scores. In addition to each type of test score, these methods use the linguistic features of dictation/read-aloud materials. The research question was answered by testing the equality between the statistically significant correlation coefficients in the chi-square tests.

5 Experimental Results and Discussion

The mean, minimum, and maximum GEP of the 50 learners were 607.7, 295, and 900, respectively, and the *SD* was 184.45.

Table 1 shows the means and SDs of the dictation and read-aloud performance scores. Table 2 shows the means and SDs of the linguistic difficulty of sentences in the text material.

Table 3 shows the correlation coefficients between the measured and observed GEPs in the cross-validation tests. Here, D&R refers to a measurement method using dictation and read-aloud, D represents a method using dictation, and R denotes the method using read-aloud. When the correlation coefficient was significantly different from zero, it was marked with an asterisk in all three types of measurement methods.

Table 4 shows the results of the chi-square tests for the equality of correlations among the three measurement methods. Bold chi-square values indicate significant differences between the correlation coefficients.

Table 3 shows the values of correlation coefficients in a descending order: D&R > D > R. Table 4 indicated the statistical significance of pairs of correlation coefficients in the descending order: D&R > D, D&R > R, and D > R. The measurement method using D&R demonstrated the strongest correlation. That is, the results suggest that D and R are complementary for measuring GEP.

The significant difference in D > R suggests that spelling is more associated with TOEIC than pronunciation. In D, learners output grapheme strings, while they output phoneme strings in R. The former needs more sophisticated language ability because the errors in spelling can be more clearly identified, and learners use the visual, auditory, and haptic (kinesthetic and tactile) senses (Dobie 1986). Hence, the correlation result, or D > R, can be considered evidence that D is more associated with TOEIC than R.

Therefore, this study suggested that GEP should be measured with a method using D&R because of the strength of correlation, that is, D&R > D > R. However, if teachers must decrease the time for test administration and/or to reduce preparation tasks for test materials, a measurement can also be developed with only D instead of using D and R.

Performance score	***n***	**Mean**	***SD***
EASE-D	750	4.22	0.77
ACC-D	750	0.44	0.19
EASE-R	750	3.03	0.91
ACC-R	750	0.95	0.06
RATE-R	750	100.66	27.39

Table 1: Descriptive statistics of the dictation and read-aloud performances

Linguistic features	***n***	**Mean**	***SD***
Sentence length	15	21.93	7.57
Mean word length	15	1.26	0.11
Number of multiple-syllable words	15	5.93	2.84
Word difficulty	15	0.26	0.11
Speech rate	15	178.44	17.41

Table 2: Descriptive statistics of the linguistic difficulty of the sentences

Measurement methods	***r***	***t***	***df***	***p***
D&R	0.80*	36.13	748	< 0.05
D	0.75*	31.17	748	< 0.05
R	0.59*	19.78	748	< 0.05

Table 3: Correlation coefficients of the three measurement methods

Measurement methods	***chi sq.***	***df***	***p***	***FDR***
D&R > D	**4.89**	1	0.03	0.05
D&R > R	**65.79**	1	< 0.02	0.02
D > R	**34.78**	1	< 0.03	0.03

Table 4: Chi-square tests for equality among the three measurement methods

6 Conclusion

This study determined which GEP-measurement method achieved the best performance. The three GEP-measurement methods were developed using dictation and/or read-aloud performance scores as well as the linguistic features of the dictation/read-aloud materials. These methods were compared respecting the measurement accuracy and ease of measurement.

The experimental results suggested that GEP should be measured with the dictation and read-aloud-based method, as the measured GEP had the strongest correlation with the observed GEP. However, if teachers must decrease testing time and/or preparation tasks for test materials, the dictation-based method can also be utilized.

Future research should examine what combinations of dictation performance (EASE-D and ACC-D) and read-aloud performances (EASE-R, ACC-R, and RATE-R) can achieve the best measurement performance. How the measurement depends on learners' GEP should also be investigated.

Acknowledgments

References

Yoav Benjamini and Yosef Hochberg.1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289-300.

Jeanne S. Chall and Harold E. Dial. 1948. Predicting Listener Understanding and Interest in Newscasts. *Educational Research Bulletin*, 27(6):141-153+168.

David Coniam. 1991. Reading Aloud Speed as a Factor in Oral Fluency and General Language Proficiency?. *Hongkong Papers in Linguistics and Language Teaching*, 14:47-69.

David Deterding. 2006. 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.

Ann B. Dobie. 1986. Orthographical Theory and Practice, or how to Teach Spelling. *Journal of Basic Writing*, 5(2): 41-48.

Educational Testing Service. 1998. *TOEIC Technical Manuel*. Educational Testing Service, Princeton: NJ.

Irving E. Fang. 1966. The Easy Listening Formula. *Journal of Broadcasting*, 11(1):63-68.

Atsushi Iino, Yukiko Yabuta, and Joel Thomas. 2011. Relationship between Criteria for Reading Aloud Evaluation and English Proficiency. *Journal of the Chubu English Language Education*, 40:159-166.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge, UK.

Patricia Irvine, Parvin Atai, and John W. Oller, Jr. 1974. Cloze, Dictation, and the Test of English as a Foreign Language. *Language Learning*, 24(2):245-252.

Masaya Kanzaki. 2015. Minimal English Test: Item Analysis and Comparison with TOEIC Scores. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 19(2):12-23.

Semin Kazazoglu. 2013. Dictation as a Language Learning Tool. *Procedia-Social and Behavioral Sciences*, 70:1338-1346.

Hideo Kiyokawa. 1990. A Formula for Predicting Listenability: The Listenability of English Language Materials 2. *Wayo Women's University Language and Literature*, 24:57-74.

Katsunori Kotani and Takehiko Yoshimi. 2021a. Predicting English Proficiency with Read-Aloud Performance and Linguistic Difficulty of Sentences. *Proceedings of International Technology, Education and Development Conference*, 1180-1185.

Katsunori Kotani and Takehiko Yoshimi. 2021b. Prediction of General ESL Proficiency Considering Learners' Dictation Performance. *The 3rd ETLTC International Conference on Information and Communications Technology, EDP Sciences, France*. https://doi.org/10.1051/shsconf/202110201003.

Satsuki Kojima and Soichi Ota. 2012. Shadowing, Dictation and Reading Aloud: Which is Effective?. *Journal of The Japan Association of College English Teachers*, 4:29-40.

Yongeun Lee. 2014. Quantifying English Fluency in Korean Speakers' Read-Aloud and Picture-Cued Storytelling Speech. *Linguistic Research*, 31(3):465-490.

Paul Leeming and Aeric Wong. 2016. Using Dictation to Measure Language Proficiency: A Rasch Analysis. *Language Testing and Assessment*, 5(2):1-25.

David Meyer, et al. 2021. e1071: *Misc Functions of the Department of Statistics (Formerly: E1071)*, TU Wien. https://cran.r-project.org/web/packages/e1071/e1071.pdf.

Jose Noijons. 1994. Testing Computer Assisted Language Testing: Towards a Checklist for CALT. *CALICO Journal*, 12(1):37-58.

John W. Oller, Jr. 1983. *Evidence for a General Language Proficiency Factor: An Expectancy Grammar*. In John W. Oller, Jr. (Ed.), Issues in Language Testing Research. Newbury House, Rowley: MA.

Anthony Stenton. 2013. 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*, 2(1):145-161.

Ruslan Suvorov and Volker Hegelheimer. 2013. *Computer-Assisted Language Testing*. In Antony J, Kunnan (Ed.), The Companion to Language Assessment, John Wiley & Sons, Hoboken: NJ.

Elvis Wagner. 2020. Duolingo English Test, Revised Version July 2019, *Language Assessment Quarterly*, 17(3):300-315.

Aeric Wong and Paul Leeming. 2014. Using Dictation to Measure Language Proficiency. *Language Education in Asia*, 5(1):160-169.

Anoushe Yazdinejad and Mitra Zeraatpishe. 2019. Investigating the Validity of Partial Dictation as a Test of Overall Language Proficiency. *International Journal of Language Testing*, 9(2):44-55.