A comparison of the validity of measurement methods of the general English proficiency by dictation and read-aloud performance

Katsunori Kotani
Kansai Gaidai University
Osaka, Japan

kkotani@kansaigaidai.ac.jp

Takehiko Yoshimi
Ryukokou University
Shiga, Japan

yoshimi@rins.ryukoku.ac.jp

Abstract

This paper compares three classroom-based measurement methods of the general proficiency of English as a second language (GEP): one measures the GEP by dictation performance; another by read-aloud performance; and the other by both dictation performance and read-aloud performance. A learner's GEP has been measured by commercially tests because the reliability and validity of the tests have been well-acknowledged, but the use of the tests has ineffective regarding test-administration cost and test materials. By contrast, measurement of dictation and/or read-aloud needs only a sound file or a text file, and a teacher of English as a second language can choose test materials according to the class contents. This study developed the three GEP-measurement methods using a learner corpus data involving GEP, dictation and read-aloud performance of 50 learners of English as a second language. The experimental results suggested that the proficiency should be measured with both dictation and read-aloud performance.

1   Introduction

In a class of English as a second language (ESL), a teacher should conduct test in a classroom, because classroom testing has benefits to grow students' general English proficiency (GEP). Classroom testing provides a teacher with valuable feedback about students' learning outcome (Roediger et al. 2011). Thus, a teacher can understand how students' GEP grows. In addition, frequent testing encourages students to study (Roediger et al. 2011). Thus, if a teacher provides a test in a classroom, students will increase learning motivation.

A classroom testing method is necessary, because commercially available tests such as Test of English as a Foreign Language (TOEFL) or Test of English for International Communication (TOEIC) have three limitations. First, test fees are expensive for classroom testing, although it can be applicable once in an academic year for confirming learning outcome or using as a placement test to determine a class appropriate for students' proficiency. Second, test administration takes a couple of hours. It is longer than a university class period (90–105 minutes in Japan). Finally, the test material is irrelevant for the ESL classes being taken by students.

A solution to these limitations is to introduce a computer-assisted language testing (Noijons 1994, Suvorov 2013). Here, a learner's GEP was measured by calculating scores for a learner's reading aloud performance and dictation performance. Dictation and read-aloud performances were taken up in this study for four reasons.

First, tests for dictation and read-aloud have been reported to demonstrate the reliability and validity as a GEP test (Kotani & Yoshimi 2018, Kotani & Yoshimi 2020). In addition, dictation and read-aloud tests have also been reported to demonstrate GEP (Irvine, Atai, and Oller, Jr. 1974; Oller 1983; Coniam 1991; Lee 2004; Iino, Yabuta and Thomas 2011; Kazazoglu 2013; Wong and Leeming 2014; Leeming and Wong 2016; Yazdinejad and Zeraatpishe 2019). Given

the close relation of GEP with dictation and read-aloud performance, dictation and read-aloud tests are used in Duolingo, commercially available testing for GEP (Wanger 2020).

Second, these tests save an ESL teacher's time and efforts to prepare and administer reading/listening comprehension questions. An ESL teacher only must prepare a sound/text file for dictation/read-aloud. Thus, an ESL teacher can use a dictation/read-aloud test as a quick check for GEP several times in a semester/in every class. In addition, a teacher can choose test materials according to the class contents.

Third, a dictation/read-aloud test is low cost for evaluation. In a dictation test, students' answers can be evaluated by comparing with reference sentences manually by an ESL teacher or automatically by a natural language processing tool to measure edit distance. In read-aloud test, the accuracy of pronunciation can also be evaluated by comparing with reference pronunciation manually by an ESL teacher or automatically by a sound recognition tool (Fu et al. 2020).

Finally, both dictation and read-aloud tasks are also effective from the pedagogical viewpoint. Kojima and Ota (2012) investigated the effect of dictation, read-aloud and shadowing by comparing test results between a pre-test and a post test of a semester. The results indicated that dictation could improve listening ability than shadowing.

Previous research can be 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 of 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 by dictation performance and read-aloud performance, respectively. Therefore, it has not been examined to what extent the measurement performance can be improved by measuring GEP based both on dictation and read-aloud performances.

The goal of this study is to find 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 with respect to the measurement accuracy, but also with respect to the easiness of measurement, that is, cost for the development of a method and the administration of the method.

The contribution of the present study includes (1) proposal of effective GEP-measurement method as a classroom-based assessment alternative to GEP tests, (2) empirical verification of GEP measurement methods, i.e., a dictation-based method, a read-aloud method, and a both dictation and read-aloud method, and (3) empirical verification of robustness of a GEP measurement method against English-language-class size of training data.

2 Related Research

2.1 GEP Measurement based on Dictation or Read-Aloud Performance

Kotani and Yoshimi (2021a) investigated the validity of dictation performance as an indicator of GEP by examining GEP measurement using non-linear regression analysis. The participants were 50 college students who dictated two texts that they were familiar with. The dependent variable was GEP in terms of TOEIC scores, and the independent variables included the dictation performance based on accuracy and the learners' subjective judgment, and linguistic features of the dictation texts. The measured GEP had a strong correlation with the observed GEP.

Kotani and Yoshimi (2021b) investigated the validity of read-aloud performance by examining GEP measurement using multiple linear regression analysis. The dependent variable was GEP in terms of TOEIC scores, and the independent variables included the read-aloud performance based on accuracy, speech rate, and the learners' subjective judgment, and linguistic features of the read-aloud texts. The measured GEP had a moderate correlation with the observed GEP.

2.2 Correlation Analysis between GEP and Dictation or Read-Aloud Performance

Previous research (Irvine, Atai, and Oller, Jr. 1974; Kazazoglu 2013; Wong and Leeming 2014; Leeming and Wong 2016; Yazdinejad and Zeraatpishe 2019) presumed that dictation performance is a good indicator of GEP, and

compared learners' dictation performance scores with GEP scores such as the TOEIC and the TOEFL. The results showed a strong correlation between the dictation performance scores and GEP test scores. Hence, previous research has succeeded in demonstrating that dictation performance indicates test scores for GEP.

Coniam (1991) investigated the validity of speech rate in reading aloud as an indicator of GEP by examining the correlation between GEP and the speech rate. The participants were 83 secondary school students who read a short dialogue aloud (the length was uncertain). The read-aloud performance was evaluated on a seven-point Likert scale and included accuracy and fluency. GEP scores constituted reading-writing and speaking-listening scores. The speech rate was moderately correlated with read-aloud performance ($r = 0.66$), reading-writing performance ($r = 0.55$), listening-speaking performance ($r = 0.60$), and GEP ($r = 0.59$).

Iino, Yabuta and Thomas (2011) investigated the validity of read-aloud performance scores as an indicator of GEP by examining the correlation between GEP and read-aloud performance. In the read-aloud tests, 80 ESL learners read aloud four short texts. The results of the read-aloud tests were evaluated by three ESL teachers on a five-point Likert scale based on accuracy of pronunciation and accent, intelligibility of meaning units, and fluency of speech rate. The read-aloud scores had moderate correlation with test scores of GEP.

Lee (2014) investigated the validity of speech rate in reading aloud as an indicator of oral GEP by examining the correlation between oral GEP and speech rate. The participants were 46 college students who read aloud a short text. Oral GEP was evaluated by three ESL teachers and native English speakers. The oral GEP includes read-aloud performance and picture-cued storytelling. The speech rate had a moderate correlation with the read-aloud-based GEP and strong correlation with storytelling-based GEP.

3   Collection of Dictation and Read-Aloud Data

3.1   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. The use of class-size training data reveals the possibility of an ESL teacher to develop GEP measurement using training data compiled in the class.

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.

3.2   Data Collection Procedures

Data instances to determine GEP 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 from a text material, and the learners' English test scores. The dictation and read-aloud data included 750 instances gathered from 50 learners' attempts to complete the text 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 4.1).

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 (explained in Section 4.2). Recording durations were collected to calculate speech rates.

Three instructions were given to the learners: 1) Each sentence could be listened or read twice if necessary; 2) Each task should be completed at a speed natural for the learner; and 3) It was forbidden to read fast or slowly, or to return and revise a sentence after moving on to the next sentence.

3.3   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 reading-

aloud 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. It should be noted that Text I failed to encompass certain sounds, such as initial and medial /z/ and syllable initial /θ/. However, Text II included these missing sounds.

3.4 General English Proficiency

Learners' 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 (Educational Testing Service 1998), and this test has no dictation and read-aloud sections.

4 Features for Regression

This study measured GEP through regression based on dictation and/or read-aloud performance scores and the linguistic features of a sentence.

4.1 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). A lower EASE-D indicated that the learners judged the dictation to be easier.

ACC-D was calculated by dividing the Levenshtein edit distance between a given reference and a transcribed sentence 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-D denoted that the learners completed the dictation more accurately.

4.2 Read-Aloud Performance

The criteria for evaluating read-aloud performance comprised three indices. They were 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).

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 (0 indicated the absence of words correctly read aloud and 1 indicated that the learner read aloud all words correctly). Learners' reading aloud is evaluated word-by-word using a binary decision (correct or incorrect pronunciation) for constructing a naïve measurement method as a cost-effective method. The measurement accuracy might be improved if the reading aloud is evaluated phone-by-phone using the multiple decision for the appropriateness, but the development cost will increase. From the viewpoint of the cost-effectiveness, the reading aloud should be evaluated manually by an ESL teacher or automatically by a speech recognition tool, but we chose an English transcriber as an evaluator. A transcriber is supposed to provide strict evaluation for learners' reading aloud due to the unfamiliarity with learners' pronunciation. When this method is practically used in an ESL class, an ESL teacher should evaluate learners' reading aloud as strictly as possible to maintain the validity.

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

4.3 Linguistic Features

In this study, linguistic features included sentence length, mean word length, number of multiple-syllable words, and word difficulty. These linguistic features were automatically derived from sentences in the text material.

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 a basic vocabulary list 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.

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

Measurement methods were developed using support vector regression, considering GEP as the dependent variable. The independent variables were the dictation performance scores (EASE-D, ACC-D), the read-aloud performance scores (EASE-R, ACC-R, RATE-R), and the linguistic features.

GEP is measured by support vector regression. Support vector regression is less explainable but more accurate than multiple regression. Thus, an active feature of a learning model is not clearly described. 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, considering one instance as test data and *n*-1 instances as training data. The training/test data included 750 instances.

Correlation analysis using *t*-test was performed between the measured and observed GEPs, where the significant threshold was set to 0.05. A statistically significant correlation was further examined using chi-square test for answering the research question, but a non-significant correlation was left out. The significance threshold was adjusted for multiple testing based on the false discovery rate (FDR) (Benjamini and Hochberg 1995).

To address the research question, three types of a measurement method were developed: one uses dictation performance scores, another read-aloud performance scores, and the other both dictation and read-aloud performance scores. In addition to each type of test scores, these methods use 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.

6 Experimental Results

6.1 Descriptive Statistics

Figure 1 shows the distribution of GEP. GEP followed a normal distribution according to the Kolmogorov-Smirnov test ($K = 0.82$, $p = 0.25$). The mean, minimum, and maximum GEP were 607.7, 295, and 900, respectively, and the *SD* was 184.45.

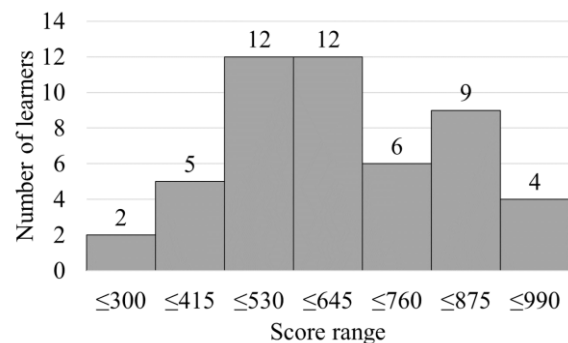Figure 1: GEP distribution

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

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

6.2 Results and Discussion

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

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

The scatterplots in Figure 2–4 show the correlations between the observed GEP and measured GEP (D&R, D, and R).
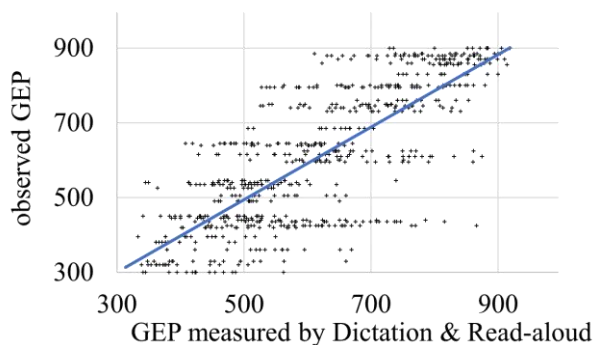

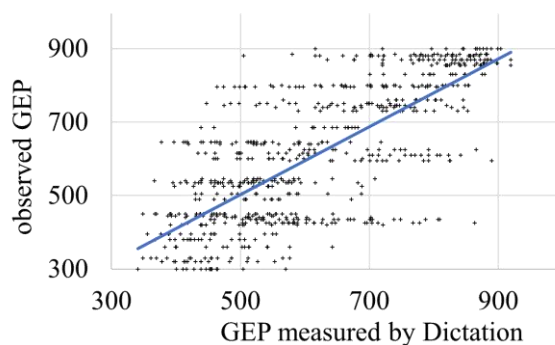Figure 2: Scatter plot of GEP measured by D&R

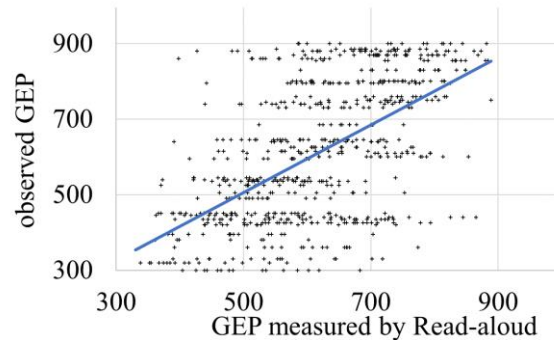Figure 3: Scatter plot of GEP measured by D

Figure 4: Scatter plot of GEP measured by R

Table 4 shows the results of the chi-square tests for equality of correlation among the three measurement methods. *FDR* refers to the significance threshold adjusted for multiple testing based on the false discovery rate. The chi-square value marked with an asterisk indicates significant differences between the correlation coefficients. The values of correlation coefficients are shown in a descending order: D&R > D > R. Table 4 indicated that statistically 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 result suggests that D and R are complementary to measure GEP.

Measurement methods	*chi-square*	*df*	*P*	*FDR*
D&R > D	4.89*	1	< 0.05	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

The significant difference in D > R can be explained in relation to association with TOEIC. Assuming that GEP can be measured with TOEIC scores, D > R indicated that TOEIC had stronger correlation with D than R. Both D and R share listening and reading comprehension skills, respectively. However, spelling in D and pronunciation in R are not examined in TOEIC. Hence, the correlation result, i.e., D > R, can be taken as a piece of evidence that spelling is more associated with TOEIC than pronunciation.

Therefore, the present study suggested that GEP should be measured with a method using D&R because of the strength of correlation, i.e., D&R > D > R. However, if an ESL teacher needs to decrease time for test administration and/or to reduce preparation tasks for test materials, a

measurement can also be developed only with D instead of using D & R.

7 Conclusion

This study answered which of the GEP-measurement methods achieved the best performance among a dictation-based method, a read-aloud method, and a dictation and read-aloud-based method. Each method was developed by a non-linear regression analysis using dictation and/or read-aloud performance scores, and the linguistic features of the dictation/read-aloud materials. These methods were compared with respect to the measurement accuracy and the easiness of measurement.

The experimental result suggested that GEP should be measured with the dictation and read-aloud-based method, because the measured GEP had the strongest correlation with the observed GEP. However, if an ESL teacher needs to decrease testing time and/or preparation tasks for test materials, the diction-based method can also be utilized.

Future research should examine what combinations of dictation performances (EASE-D and ACC-D) and read-aloud performances (EASE-R, ACC-R, and RATE-R) can achieve the best measurement performance. It should also investigate how the measurement is dependent on learners' GEP.

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.

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.

Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito. 2020. Automatic Assessment of English Proficiency for Japanese Learners without Reference Sentences based on Deep Neural Network Acoustic Models. Speech Communication, 116:86–97.

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, pages 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.

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, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang, Chih-Chen Lin. 2021. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

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, pages 3–10, Newbury House, Rowley: MA.

Henry L. Roediger III, Adam L. Putnam, and Megan A. Smith. 2011. Ten Benefits of Testing and their Applications to Educational Practice. In Jose P. Mestre and Brian H. Ross (Eds.), Psychology of Learning and Motivation: Cognition in Education, 55: 1–36, San Diego, CA: Elsevier Academic Press.

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.