Designing and Validating an English Speaking General Proficiency E-Test

Dr.Marwa Ahmed Refat Naeem
Lecturer of TEFL
Faculty of Education, Kafr El-Sheikh University

Abstract

English Speaking General proficiency Test is an e-test that was designed and applied to a wide range of EFL speakers: three hundred and thirty participants. The test was meant to be a valid and reliable tool that could be used as a trustful basis for estimating speaking proficiency. Seven speaking sub-skills were assessed for each testee: pronunciation, intonation and stress, vocabulary, grammar, cohesion, content and fluency. There were a number of alternatives for test delivery and administration: at language labs, on personal cell phones or on Google Drive or YouTube. Assessment rubrics were also designed by the researcher so as to score the measured speaking sub-skills. Test reliability was established by calculating Alpha Coefficient. Besides, criterion-referenced validity was statistically calculated by comparing the participants' scores on the current test with their scores on an ILETS speaking test sample. Both reliability and validity were calculated by using the SPSS 11.0 program. Results showed that the current form of the test was reliable and valid for testing English speaking general proficiency.

Key Words: Language Testing, Computerized Tests, Speaking Assessment, Proficiency Tests, Evaluation.

Introduction

Speaking is a basic language skill. It is a productive oral-mode skill in the sense that the linguistic material is produced orally by the speaker himself. However, this skill is usually neglected in most EFL and ESP courses as well as classes. Most formal exams in different levels and grades concentrate on testing reading and writing rather than listening and speaking.

The theory and practice of testing second or foreign language speaking general proficiency is considered to be the youngest sub-field of language testing. It was the Second World War that constituted a turning point in the interest in and focus on testing speaking. The reason behind that was the need for

revolutionizing teaching and testing speaking so as to serve political and military goals. As for testing speaking in educational contexts, the earliest testing system was adopted by many universities and schools so as to certify the proficiency of bilingual teachers and academics in the 1970s.

Questioning the value of the TOEFL iBT speaking section as an indicator of university students' academic oral ability, Ockey et al. (2015) made a study on 222 English university majors. They compared TOEFL iBT speaking scores to performances on a group oral discussion, picture and graph description, prepared oral presentation tasks. Pronunciation, fluency, grammar, vocabulary, interactional competence, descriptive skill, delivery skill, and question answering were the speaking sub-skills concerned in the comparison. Statistically, they used Pearson product-moment correlation between scores on the speaking section of TOEFL iBT and some university speaking tasks. TOEFL iBT speaking scores were found good overall indicators of academic oral ability. Although they were better measures of pronunciation, fluency, vocabulary and grammar, they were not that efficient in measuring interactional competence, descriptive skill and presentation delivery skill.

Mianto (2014) discussed the difficulties in testing speaking performance. She also elaborated rubrics as a tool to grade speaking tests. Three types of speaking tests: monologue, dialogue and multilogue were compared in terms of their different characteristics and purposes. Furthermore, Mianto depicted two main difficulties in testing speaking. She illustrated that speaking involved a combination of skills that might have no correlation with each other, and which did not lend themselves well to objective testing. Besides, there were many language features in speaking skill that became influence factors of scoring. In conclusion, she recommended using rubrics for testing speaking mentioning their merits.

On the same track, Hughes (2013) discussed the difficulties in the way of teaching and testing speaking. She asserted the overlapping nature of speaking with a considerable number of

other areas and disciplines as a central difficulty of developing this skill.

There was a reassertion of the importance of speaking ability after the expansion in the constructs underlying second / foreign exams. It was argued that it could be taken for granted that the ability to express oneself in writing was correlated with a similar ability to speak the language fluently. This increased the interest in teaching and testing speaking (Weir et al., 2013).

As for the development of testing speaking skills, Weir et al. (2013) added that testing speaking broadened from the relatively narrow conceptualization of speaking as pronunciation accuracy, at the beginning of the 20th century, to a communicative and later an interactional definition of the construct. Speaking could be measured by oral tests that included a range of tasks allowing the display of communicative language ability in a variety of contexts. They noted that the widespread use of computer-delivered and computer-scored oral tests had succeeded in measuring the core features of speech. These core features were generated in monologic tasks and features of language.

Examining six best known English proficiency tests; namely: Cambridge ESOL Exams, TOEFL, ILETS, Trinity College London Exams, Pearson Tests of English and the International Test of English Proficiency; Varela and Palacios (2013) focused on the different tasks and assessment criteria for oral production skills. They argued that spoken skills constituted an important part of general English examinations. Oral skills were assessed through a variety of tasks that included personal interviews, photo descriptions, topic discussions and role play. It was also noted that there was a preference of computer-based speaking examinations over face-to-face examinations in short question-answer tasks, whereas face-to-face examinations were more effective in combining such tasks as personal interviews and topic discussions. Besides, speaking was often assessed on the basis of four trait range: accuracy, fluency, interaction and coherence. However, Varela and Palacios (2013) concluded that

there were points of weakness in the investigated English speaking tests. The first defect they found was the ambiguity of the grammatical features under consideration. In addition to this, there was a lack of updating or adapting mechanisms in the guides of these tests so as to keep up with new developments in English language teaching methodologies. A third defect was represented in the lack of focus on the "human" element – those who took the test and those who scored it – during test design.

Paker and Höl (2012) explored the attitudes and perceptions of the students and instructors towards the speaking test at a School of Foreign Languages. Their sample included 210 students and 32 instructors. Final results indicated that most of the students had no experience of any speaking test before, and therefore, they had higher anxiety during the test. Among students, the speaking test was regarded as the most difficult test when compared to the testing of other language skills. Moreover, students pointed out that they could not express themselves adequately during the test, and claimed that they needed to have more oral practice in the classroom. As for instructors, it was emphasized that the speaking test was the most difficult one to apply and assess; however, the scale and rubrics were adequate enough to assess the students' oral performance.

Overviewing the academic literature on face-to-face and computer-based assessment of speaking and exploring the test features of these two different test modes, Galaczi (2010) concluded that the main advantage of computer-delivered and computer-scored speaking tests was their convenience and standardization of delivery, which enhanced their reliability and practicality. However, face-to-face speaking tests and the involvement of human interviewers and raters introduced a broader test construct, since interaction became an integral part of the test, and so learners' interactional competence could be tapped into. She argued that there was not just one way of testing speaking, or one 'best' way to do it. Thus, language testers should choose from a range of useful formats which aid in eliciting and assessing speaking skills, from fully automated speaking tests to

ones involving human interviewers and raters. This recommendation inspired the integrative design of the current test.

In the same vein, Qian (2009) was preoccupied with the debates over the appropriateness of two different testing modes, namely, (a) face-to-face, or direct, testing, and (b) person-to-machine, or semi-direct, testing. He found the results of the previous research conducted in this area mixed and confusing. His investigation was carried out in the context of a university setting in Hong Kong and compared the popularity of both testing modes by analyzing reactions and perceptions of a group of test takers who had sat for both test modes.

The results indicated that although a large proportion of the participants had no particular preference in terms of the testing mode, the number of participants who strongly favored direct testing far exceeded the number strongly favoring semi-direct testing. The participants' main reason cited for disliking semi-direct testing was its inability for the examiner and examinee to interact during the test, which appeared to have created a psychological barrier for the test taker.

Opposing the common belief that tests of spoken language ability are the most difficult, O'Sullivan (2008) highlighted the recent improvements in designing and evaluating speaking tests. He depicted the different characteristics of the speaking context, the speaking test taker and the speaking tester. Afterwards, he thoroughly discussed and compared between the holistic rating and the analytical rating scales of speaking skills. His paper concluded that there was a very high correlation between a test taker's holistic score and the total analytic score. This reflected the findings of a number of studies in which both analytic and holistic scores were given, and certainly suggested that both scales offered very similar outcomes.

Reviewing recent trends in the conceptualizations and formats of English proficiency tests, Cumming (2007) focused on construct validation, consistency and innovations in the media of

test administration including various forms of computer and other technological adaptations of a number of common tests such as TOEFL. Cumming's analysis and criticism were enlightening in guiding the current test design.

According to Loma (2004), speaking assessment was a cycle process that involved four stages. The first stage was to recognize a need for speaking assessment. The second stage was planning and developing the assessment tasks and criteria. Two interactive stages followed to complete the cycle: test administration and test rating or evaluation. This cycle was followed by the researcher in designing and implementing the test in hand.

Handling speaking sub-skills and assessment, Florez (1999) pointed out that speaking sub-skills included language functions (the patterns that tend to occur in certain discourse situations); linguistic competence (producing specific points of language such as grammar، pronunciation or vocabulary) and sociolinguistic competence (understanding when, why and in what way to produce language). She also argued that speaking assessments could take many forms, from oral sections of standardized tests to authentic assessments. An assertion was on the criteria that should be clearly defined and understandable to both the testers and testees.

Describing how speaking tests were conducted and evaluated, Adams (1979) pointed out that a speaking test began with simple social formulae such as introductions, comments on the weather or other ice-breaking questions. The testee's response to these questions identified the preliminary ceiling of the course of the rest of the test. A testee was commonly asked to talk about himself, his family and his work. He might be asked to play a role or give street directions. A testee's adequate coping to preliminaries led the tester to move to natural conversations on autobiographical and professional topics. In respect of evaluating speaking tests, Adams (1979) provided a checklist that consisted of five sub-skills; namely: accent, grammar, vocabulary, fluency and comprehension. She explained that accent was divided into

pronunciation and intonation; and grammar included morphology and syntax. The gradual difficulty arrangement of test questions that was followed by Adams (1979) as well as the checklist technique in evaluation was of great help for designing and scoring the current test.

Defining the Construct of Speaking

Testing speaking required defining the construct through answering such questions as: what is speaking? And, what constitutes speaking abilities? A construct could not be observed directly. However, it could be defined in terms of the observable behaviours of interest in a particular learning context. Speaking as a construct had to be associated with acts that could be observed and, later, could be scored. In a nutshell, all the sub-skills involved in the process of testing speaking had to be defined operationally. It was argued that to test speaking, a tester should 'pick and mix' so as to define the construct. A rationale and an empirical evidence should be provided to support the mix in the light of the test purposes (Fulcher, 2014).

On the track, Long and Doughty (2011) argued that there were two necessary parameters to define the construct of speaking: the repertoire and the explanatory conditions. The repertoire referred to the range of features and combination of features which speaking included. The explanatory conditions included the range of basic and socio-psychological conditions that explain the occurrence of the speaking features.

Reviewing related literature and previous speaking tests, it was found that the basic speaking sub-skills that could be operationally measured were:

1.Pronunciation

Pronunciation refers to the production of sounds to make meaning. It includes attention to the particular sounds of a language (language segments). Regarding pronunciation to be a set of habits to produce sounds, it was thought that learning to pronounce a second language meant building up new

pronunciation habits and overcoming the bias of the first language (Gilakjani, 2012).

2.Intonation and Stress

Intonation is defined as the melody of speech. On testing intonation, a tester observes how the pitch of voice rises and falls and how the testee uses pitch variations to convey meaning. If there was no intonation, speech would be monotonous. Intonation has four functions: (a) the attitudinal function, which expresses the speaker's attitudes and emotions; (b) the grammatical function, which identifies speech structure; (c) the informational function, which distinguishes new pieces of information in an utterance and (d) the cohesive function, which signals the contrast or the coherence of clause sequences (Wells, 2006).

Concerning stress, it is the combination of loudness, pitch and duration. English is a stress language: stress is an important component of each word. Stress can distinguish word meaning (*'billow* and *be'low*) or identify its part of speech ('*import* "noun" and *im'port* "verb") Wells (2006).

3.Vocabulary

Vocabulary has to do with the knowledge of words. Milton (2009) illustrated that there was no clear-cut definition of a "word". In tests that measured vocabulary knowledge, testers defined the term 'vocabulary' according to the circumstances and learners' characteristics. Thus, the current author in this e-test defined vocabulary as the count of the word repertoire that a testee used to answer each question.

4.Grammar

In general, grammar involves two branches: morphology and syntax. Morphology is the structure of words. Syntax is the structure of sentences. Testing grammar has a number of levels. In an advanced level, grammar is tested to check whether its rules generate the expressions a speaker wants to say and do not generate the ones he does not want (Larson, 2010). In this test,

the speaker's ability to use correct word and sentence structures identifies his / her level at the grammatical skill.

5.Cohesion

The function of cohesion is to make a spoken passage of any length form a unified whole. This means that a cohesive spoken utterance is not just a collection of separate, unrelated sentences. More specifically, cohesion occurs where the interpretation of some element in the spoken utterance is dependent on that of another. This creates an integrated spoken utterance (Halliday & Hasan, 2014). Cohesion is achieved through the choice of conjunctions and connectors. It includes the physical 'internal ties' of discourse (Renkema, 2009).

6.Content

The sub-skill of content refers to the speaker's ability to state related, enriched or expanded information about each stimulus. On the excellent level of content, a speaker is capable of giving a creative response to the stimulus.

7.Fluency

Fluency can be defined as rapid، smooth, accurate, lucid, and efficient translation of thought into the target language. In a simple simile, fluency is seen as the fluidity and automaticity of speech. The aspects of fluency include smoothness, confidence and accurate expression and rate (Sidimanjana et al., 2014).

Purpose of the Test

This test was designed to be a valid and reliable tool to avoid the shortcomings of the well-known speaking tests and to measure English speaking general proficiency. It was not limited to a certain content. Through tasks and elicitation techniques, the speaking test in hand collected evidence in a systematic way to support an inference about the speaking construct and its defining sub-skills.

Need for the Test

Measuring English speaking general proficiency may be a target for:

Academic EFL Departments

Some specialized academic departments in universities think of a criterion upon which candidates are accepted or rejected. The current test can be used as a selection test to decide on the acceptance or rejection of students into a particular program.

Public and Private Institutions

Airports, tourism companies, banks, diplomatic institutions, language centres and the like may include posts that entail an adequate level of oral-mode skills. Thus, the current speaking test may serve to judge such job applicants' communicative skills.

EFL Researchers

Researchers who are preoccupied with speaking skills may employ the current test as a tool in their studies. Using reliable and valid tests saves time and effort for researchers and logically renders trusted results.

EFL Teachers

Teachers are always preoccupied with evaluation. Evaluating students' skills is an essential task in teaching profession. Thus, the current test may be beneficial for EFL teachers in such contexts when measuring speaking general proficiency is required.

Individuals

Whoever studies English as a Foreign Language may need to test his speaking general proficiency. This test – along with its scoring rubrics – can aid learners to self-assess their speaking skills.

Participants

The current test was taken by 330 participants. Ten graduates represented the first category of test takers. Their ages ranged from 24 to 25 years old. The second category of test participants was 170 General Secondary Certificate students. They are about 16 years old. The third category consisted of 150 participants of university undergraduates. They were 19 to 20 years old. The heterogeneity of the sample was intentional so

that the test validity and reliability could be trustfully calculated. Student participants were told that the aim of the test had nothing to do with their formal school or faculty evaluation. The test aim was pointed out to participants who took part in the test voluntarily.

Test Blueprint

The *English Speaking General proficiency Test* is a computerized evaluation tool designed by the researcher so as to measure speaking proficiency on seven speaking sub-skills; namely: pronunciation, intonation and stress, vocabulary, grammar, cohesion, content and fluency. Microsoft PowerPoint 2010 is used to design the test screens. The test includes twenty eight screens starting and ending with covers. The general instructions of the test are displayed on the screens (2) through (5). They point out the aim of the test and its allotted time. Moreover, these general instructions give the gist of the test structure and when and how to answer its questions. On Screen (6), the testee is instructed to press the "Start Button" to begin doing the test. Screen (7) displays the instructions of Part I. Part I questions are displayed on six screens. This part contains two direct questions and takes five minutes. Afterwards, Screen (14) appears automatically to give the testee instructions on Part II. In this part, the testee is required to answer three questions. A testee looks at some pictures on the screen and tries to describe or comment on them. The time allotted to this part is six minutes. The instructions of Part III appear on Screen (24). This part consists of only one question. It requires a testee to give a short talk on a certain topic. The allowed time for this part is four minutes. The last screen displays the end cover of the test.

Test Administration

On delivering the test, two formats were available to participants: a video format and a PowerPoint show file format. Another choice was given to participants. They could do the test at school or faculty labs, do it on their personal cell phones or do it on the Internet at home as the test was uploaded at Google

Drive and Youtube. The availability of such alternatives made students feel at ease and encouraged to take the test.

A number of preservice teachers – who had field training at secondary schools – were instructed on how to administer the test. They took the role of testers and applied the test to secondary school students, whereas the researcher tested college graduates and undergraduates. In this test, a tester was just a guide or counselor who instructed students on how to do the test and the different alternatives to record and deliver the answer.

Scoring Rubrics

A scoring sheet was designed by the researcher to calculate the score for each testee. An analytic scoring rubric – designed by the researcher – was used to assess the target speaking sub-skills. These sub-skills were pronunciation, intonation and stress, vocabulary, grammar, cohesion, content and fluency. Each sub-skill was given a score on a scale ranged from 1 to 4. There was a detailed specification of the criteria which were required to give a certain score. Getting 1 on pronunciation – for instance – meant that the testee had a poor ability in pronouncing sounds and words. A score of 2 was interpreted as "Fair". Getting 3 meant that the testee had a good deal of the sub-skill. Finally, a score of 4 indicated the excellent mastery of the sub-skill. The sum total of the entire test was 168 scores.

Test Worthiness

According to Neukrug & Fawcett (2014), test worthiness referred to how good a test was. It encompassed an involved and objective analysis of four critical features: reliability, validity, cross-cultural fairness and practicality.

The test was tried out so as to calculate test reliability and validity. Three hundred and thirty participants took the test and the data collected were analyzed statistically using the SPSS 11.0 program.

Reliability

The first administration of the test was in September 2014. Three months later, participants were retested. Participants' responses were scored by two professional raters other than the researcher. Then, inter- rater average scores were calculated. Afterwards, reliability was calculated by finding Cronbach's Alpha Coefficient. Test-retest data were statistically processed by SPSS 11.0. It was found that Alpha Coefficient equaled .9. According to Gliem & Gliem (2003), an Alpha Coefficient that is > .9 indicates an excellent consistency. Consequently, it was concluded that the test was reliable.

Validity

Establishing test validity on a statistical basis, the participants took a model of ILETS speaking test. They were assessed according to ILETS Assessment Criteria. Criterion-referenced validity was figured by finding Pearson's Correlation Coefficient between the participants' total scores on the current test and their total scores on the ILETS speaking test.

Before finding the correlation coefficient, a necessary step was taken to test the normality of data as a prerequisite for calculating the correlation. In this concern, Shapiro-Wilk Test of normality was carried out by the SPSS 11.0 program. The following table shows the normality test results:

Table 1: Shapiro-Wilk Test of Normality

	Statistics	df	Significance
Current Speaking Test	.968	330	.425
ILETS	.946	330	.100

The significance of data of the current speaking test was .425, whereas; the significance of data of the ILETS Test was .100. Since the significance of data was greater than .05, data were normal and valid for calculating the correlation coefficient (Rovai et al., 2013).

Using the SPSS 11.0 program, correlation between participants' scores on the current speaking e-test and their scores on ILETS was estimated using Pearson correlation

coefficient. Results indicated that the correlation coefficient was .858. Consequently, it could be deduced that the scores on both tests were strongly correlated. Accordingly, the current *English Speaking General proficiency Test* was statistically valid.

Cross-cultural Fairness

The items of this test were free of cultural bias. Whether the testee was an Egyptian or an American, nothing would affect the way he understood or answered the questions. A panel of jurors in the area of ELT was consulted to guarantee that each item of the test was scrutinized for readability before administration. Questions about personal details, hobbies and suffered problems were all so common for both native and non-native English speakers alike. The three provided photos in Part II were also free from cultural connotations.

Practicality

Test practicality involved a number of factors. The ease of understanding and administering the test was one of the indicators of its practicality. The testers had a variety of techniques to administer the test: he could administer it at school, or send it to the students' e-mails, Facebook accounts … etc. A testee had also a number of alternatives: he could take the test at school or at home. The voice file that contained the answer could be recorded on a computer set or simply on a mobile phone. It could be delivered in an e-mail, on a flash memory, or simply, via a Bluetooth connection on mobiles.

The number of test takers did not affect the ease of its administration. On the contrary, the different available alternatives through which the test could be delivered allowed the testers to evaluate a relatively large sample: 330 testees.

Another point that suggested the practicality of the current test was the obvious printed design of its scoring rubrics. The set criteria for scoring each speaking sub-skill on each item saved time and effort during the evaluation and result interpretation phases.

Cost was a crucial factor of the test practicality. The design of the test screens required no complicated or professional programs. Similarly, all methods of test delivery were done for no cost at all. Free Internet and cell phone utilities were best used in this concern.

Conclusion

Designing an English Speaking General Proficiency Test is an attempt to offer a simple, reliable, valid and easy to handle tool to evaluate one of the main language skills. It endeavours to overcome the unwelcome features of the well-known speaking tests. Such features, as summed up by Abedi (2010), included not being sensitive enough to the needs of some subgroups of testees and having some unrelated variables to the focal measurement construct (e.g., unnecessary linguistic complexity and cultural biases in construction of items). Consequently, these features could affect the quality of high-stakes assessments for non-native English speakers. Modern technology is employed in the design and the delivery of the test so as to serve the purposes of being up-to-date, of being familiar to participants and of being easy to use. The researcher is looking forward to developing the current form of the newborn test and its scoring rubrics. Further research is invited to evaluate and improve the test. Other techniques for testing speaking and other speaking sub-skills to be tested may be also investigated.

References

Abedi, J. (2010). *Performance assessments for English language learners.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Adams, M. L. (1979). *Testing kit: The measurement of speaking and reading proficiency in a foreign language.* Washington: Foreign Service (Department of State).

Cumming, A. (2007). New directions in testing English language proficiency for university entrance, In *International Handbook of English Language Teaching*, 473 – 485. New York: Springer.

Florez, M. A. (1999). *Improving adult English language learners' speaking skills*. Washington: Center for Applied Linguistics.

Fulcher, G. (2014). *Testing second language speaking*. New York: Routledge.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: challenges and opportunities. *Computer-Based Assessment (CBA) of Foreign Language Speaking Skills.* Luxembourg: Publications Office of the European Union, 29 – 52.

Gilakjani, A. P. (2012). The significance of pronunciation, in *English Language Teaching*, 5 (4), 96.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's Alpha reliability coefficient for Likert-type scales. Presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH, October 8-10, 2003, 82 – 88.

Halliday, M., & Hasan, R. (2014). *Cohesion in English*. New York: Routledge.

Hughes, R. (2013). *Teaching and researching: Speaking*. New York: Routledge.

Larson, R. K. (2010). *Grammar as science*. Cambridge: MIT Press.

Long, M. H., & Doughty, C. J. (2011). *The handbook of language teaching*. New Jersey: John Wiley & Sons.

Loma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Mianto, E. (2014). Using rubrics to test students' performance in speaking. Retrieved from http://www.academia.edu/2205164.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Neukrug, E. & Fawcett, R. (2014). *Essentials of testing and assessment: A practical guide for counselors, social workers, and psychologists*. (3rd ed.). Kentucky: Cengage Learning.

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability

components for Japanese university students. *Language Testing*, 32(1), 39-62.

O'Sullivan, B. (2008). Notes on assessing speaking. Cornell University – Language Resource Center. Retrived from http://lrc. cornell. edu/events/past/2008-2009/papers08/osull1.pdf (10/01/2009).

Paker, T., & Höl, D. (2012). Attitudes and perceptions of the students and instructors towards testing speaking communicatively. *Pamukkale University Journal of Education*, 32 (2), 13 – 24.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, 6 (2), 113 – 125.

Renkema, J. (2009). *The texture of discourse: Towards an outline of connectivity theory*. Amsterdam: John Benjamins Publishing.

Rovai, A. P., Baker, J. D., & Ponton, M. K. (2013). *Social science research design and statistics: A practitioner's guide to research methods and IBM SPSS*. Virginia: Watertree Press LLC.

Sidimanjana, V. A., Marbun, R., & Rosnija, E. (2014). An analysis of student's fluency in playing drama "sherlock dolmes". *Jurnal Pendidikan dan Pembelajaran*, 3 (10), 1–14.

Varela, L. R., & Palacios, I. M. (2013). How are spoken skills assessed in proficiency tests of general English as a Foreign Language? A preliminary survey. *International Journal of English Studies*, 13 (2), 53 – 68.

Weir, C. J., Vidaković, I., & Galaczi, E. D. (2013). *Measured constructs: A history of Cambridge English examinations, 1913-2012***.** Cambridge: Cambridge University Press.

Wells, J. C. (2006). *English intonation PB and audio CD: An introduction*. Cambridge: Cambridge University Press.