Developing a Semi-Direct Speaking Test for Fourth Graders Using Video Conferencing

Yuditha Putri Wiwaharini
(yudithaaputri@yahoo.com)
Nation Star Academy School
Surabaya, Indonesia
&
Bartholomeus Budiyono
(bartholomeusbudiyono@gmail.com)
English Education Department
Graduate School
Widya Mandala Surabaya Catholic University
Surabaya, Indonesia

Article History

Received: 12-03-2021
Reviewed: 03-03-2022
Revised: 17-05-2022
Accepted: 30-05-2022

Keywords:
COVID-19 testing era; Semi-direct speaking test; Speaking test; Test development; Video Conferencing

DOI:
https://doi.org/10.33508/bw.v10i1.3071

Abstract

This study aimed to develop alternative English-speaking testing to be used during the COVID-19 pandemic. A semi-direct speaking test for 4 graders was the final product of this study. Cambridge curriculum with the ESL framework was used to formulate the test. It was designed using steps suggested by Bachman and Palmer. It also has been reviewed by an expert and a trial group. It asked each test-taker to tell a story using provided picture series, connectors, and past verbs within 5 minutes. A minute was given for them to study the picture series; the rest was to do the test. It was delivered by using video conferencing called Zoom. The result showed that the test developed was a valid, reliable, practical, and authentic measurement. Its reliability was proved by test-retest and interrater results. Its validity, practicality, and authenticity were proved by providing an expert judgement collected through questionnaires.

Introduction

Being able to get a good and proper education is important for all human beings. A good education leads humans to get a life improvement in the future. Unfortunately, in 2020, the world's education system needs to change its overall practice because of the COVID-19 virus spread. All education sectors are forced to close down because of the COVID-19 pandemic. According to COVID-19 Assessing the impact on the education sector and looking ahead (2020), during the pandemic phase, people are not allowed to go outside and do activities like what they have been doing so far. This

situation also happens in the education sector. Educators and students are not allowed to come to school and do the teaching-learning activities as usual. As a result, teaching-learning activities should be done online. Teachers and students have to follow governments' new regulations by conducting online classes as a replacement for not being able to come to offline schools. Many kinds of platforms and applications have been developed to support teachers in this situation. They are varied from ones with the simplest features to the ones with more complex features. Technically, schools can choose the platforms and applications to be used in their learning activity depending on their needs and preferences. Some schools, including in Indonesia, use additional applications like video conferencing to replace face-to-face meetings. According to Lee (2020), Video conferencing is a helpful tool to help people interact face to face without meeting in real life.

Having this sophisticated tool to support online learning does not mean that online learning is problem-free. This new situation caused many problems for all parties, especially teachers and students who take part in the online learning situation (Guangul, Suhail, Khalit & Khidhir, 2020). Many factors can be the barrier for teachers to teach remotely. Starting from the lack of strong internet connection, sophisticated gadgets, proper equipment to teach, assessment quality, and the skill to use the combination of all the above (Hasan & Khan,2020; Setyawan & Aryati Prasetyarini, 2020). Aboagye, Yawson, & Appiah (2021) confirmed that most of the learners were not ready with online learning. They added that the learners felt uncomfortable joining online learning because it was too indirect and personal for them. One of the most difficult problems to solve is the testing or assessment system during this online learning. Since teachers and students cannot meet in person, not all types and approaches of testing can be implemented. According to COVID 19 and higher education: Today and Tomorrow (2020), during this pandemic, teachers are puzzled about how to assess or test their students from distance. One of the most difficult skills to be tested during online learning is the students' production skills, especially speaking. This statement is confirmed by (Diana, 2021; Djafar, 2020). They found that most speaking testing during the COVID-19 pandemic was disturbed and became more difficult to deliver. As a result, most schools limit their assessment or testing by only giving home projects or testing for their students.

This situation is what happens in the first writer's teaching place. Most of her fellow teachers only asked their students to submit projects to be taken as their daily scores including mid and final term exams. The scores taken from those home-projects cannot be taken for granted because teachers cannot control that those assignments are done by the students individually without being helped in any way. Moreover, Haynie III (2003) found that students who only have taken home assignment treatments tended to be outscored, but in reality, they may not have a deeper understanding about the topics discussed. The need to adopt and integrate a traditional teaching-learning approach to real online learning is highly recommended (Krishan, Ching, Ramalingam, Maruthai, Kandasamy,De Mello & Ling, 2020). Therefore, many experts provided an alternative solution for this testing problem. One of which is known as a semi-direct speaking test. According to Larson (1984), semi-direct speaking test is an alternative

testing that can be used in certain circumstances and needs. It has the value of energy, cost, and time efficiency. In addition to that Guangul, Suhail, Khalit & Khidhir (2020) suggested changing the regular paper-based testing into something applicable such as online presentation, demonstrations, and reports making. Their findings also showed that 68% of the respondents prefer project-based testing. Online presentation was concluded as one of the most preferable assessments during remote learning. For all of the reasons mentioned above, this study aimed to develop reliable, valid, practical, and authentic testing to be used to test young learners' speaking ability and to answer the following research questions. 1.To what extent is the test reliable?

2.To what extent is the test valid?

3.To what extent is the test practical?

4.To what extent is the test authentic?

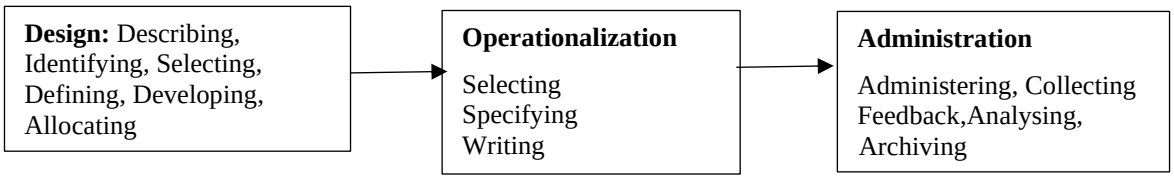
Literature Review

To achieve the above's objectives, the writers use three relevant theories to guide them. They present elaborations, examples, and deeper explanations to help them understand better. The first theory is speaking. It is taken from (Huebner, 1960). He described speaking as the main communication skill to possess. It is very important to have such a skill to interact with each other. The second theory is about semi-direct speaking test development (Bachman & Palmer, 1996). Their steps in developing a good test were taken and adapted to the current COVID-19 learning condition before applying. In their book, they defined test development as the process of making a test in detail. It starts with making the concept and design following their test-takers and curriculum used. The last theory is about video conferencing use in a test. It is taken from (Wang, 2004). He explained that video conferencing is a desktop and studio-based conferencing which is familiar to students in nowadays era. There, he provided some tips to be used to further researchers in using video conferencing in a test.

Methods

This study was a test development study. It aimed to produce a reliable, valid, practical, and authentic speaking testing to be used during the COVID-19 pandemic. The writer used a test making framework and steps adapted from (Bachman and Palmer, 1996). The following is the figure.

Figure 1. Steps of Designing a Test



They made three speaking test drafts in total before presenting their final product. The drafts and final product have been reviewed by an expert who is a Cambridge Curriculum head in a primary school taking part in this study. She reviewed the test from its deeper features matched with the curriculum used and objective set. It also has been reviewed by a similar group of targeted-test takers to see its language use and

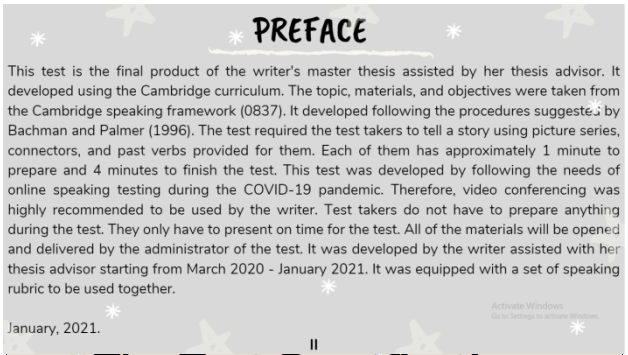
instructions from their perspectives. After getting this feedback, the writer revised the drafts until it became the final product. The test's drafts and final product were administered through a video conferencing application called Zoom. The speaking test required test-takers to be able to tell a story based on picture series given to them using vocabularies, connectors and past verbs provided. The requirements were chosen from the curriculum and framework used at the moment. The topic and type of speaking testing were chosen from the recent book chapters that targeted test-takers discussed at the moment. In scoring their performance, the writers have also developed a rubric as a set of test. It supposed to be used along with the developed test. It was made considering the criteria provided by (Nunan, 1999). She took pronunciation, task, vocabulary, and grammar to be added to her rubric. To answer the first research question, which was to what extend the test reliable, was answered by providing statistical proof of the test's test-retest and interrater reliability. The second, third and fourth research questions, which were to what extend is the test valid, practical, and authentic were answered by providing an expert judgement about the test. It was taken in the form of four Likert scale questionnaires. The expert used in this study was the Cambridge curriculum head of the school taking part in this study at the moment.

Results and Discussion

The final product of the speaking test used in this study has been revised several times following the Cambridge curriculum head and try out group feedback. They covered the use of the picture, language, instructions, guidelines to the template chosen for the test. The following was the final product of the speaking test.

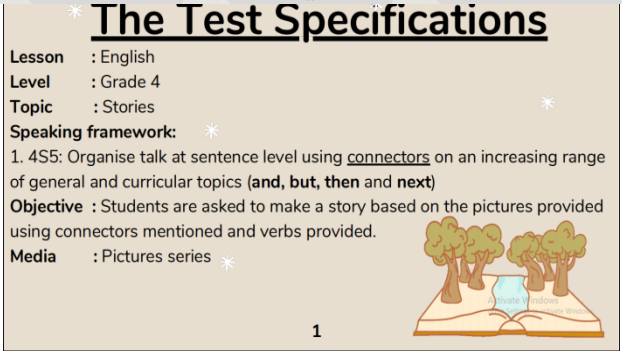
Picture 1. Test's Cover

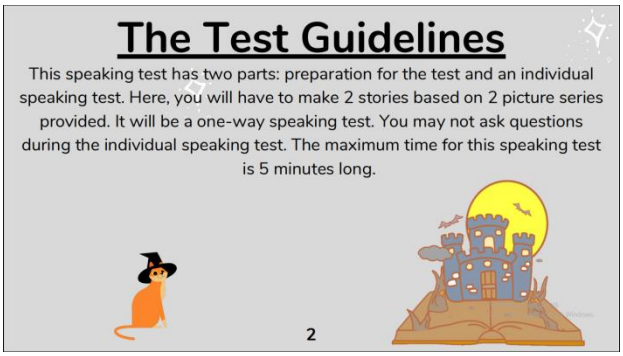
Picture 2. Test's Preface



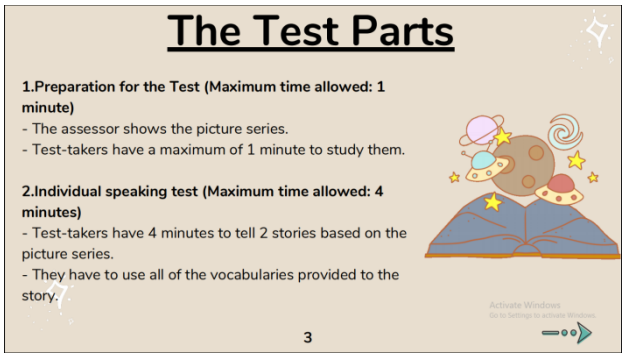
Picture 5. Test's Guidelines

Picture 6. Test's Parts

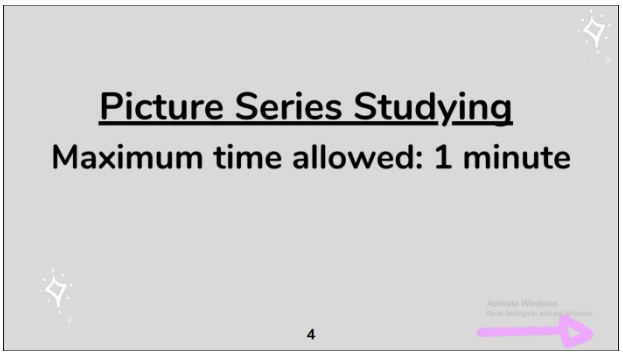




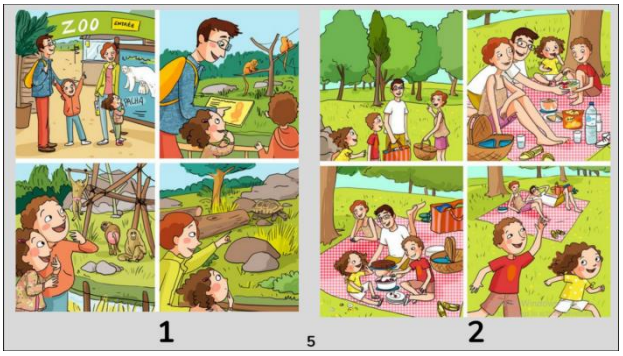
Picture 7. Picture Series Studying



Picture 8. Picture Series Showing

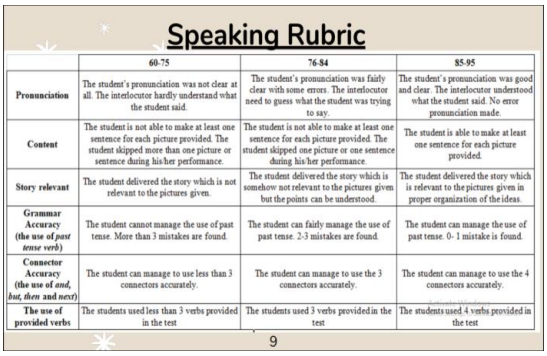


Picture 9. Test 1

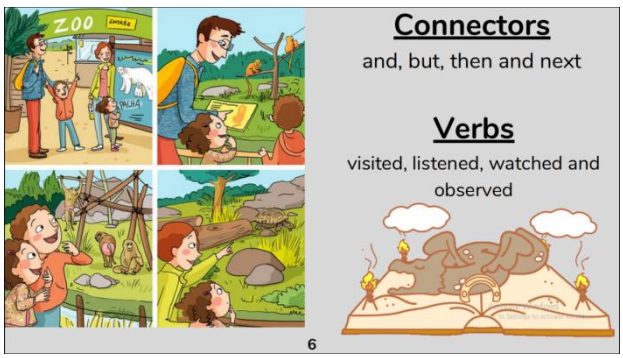


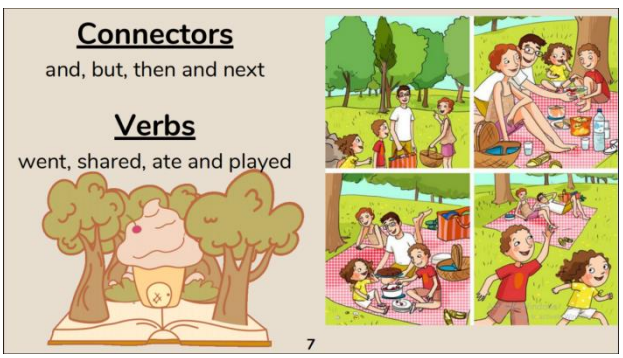
Picture 10. Test 2

Picture 11. Closing Slide



Picture 12. Speaking Rubric





This semi-direct speaking test was the final product of this study. It developed using steps adapted from (Bachman and Palmer, 1996). The framework used to design its objective was taken from Cambridge curriculum for English as a second language

with the code of 0837. Cambridge Global English for Cambridge primary, English as a second language textbook, written by Jane Boylan and Claire Medwell, published by Cambridge university press was used to choose the test's topic and materials. The vocabularies provided in the test were taken from the test-takers current discussed chapters which were chapter 5; getting around and chapter 6; school lunch. The test task asked test takers to tell a story using picture series, connectors and past verbs given to them. It was formulated following the speaking framework chosen with the code of 4S5. This test layout was made using www.canva.com. It was a free editing website that can be freely used by users around the world. Picture series used in the test were taken from www.pinterest.com. The writers have put the specific links on the references. There were two parts of the test that test takers have to follow. The first one called preparation. They were given approximately 1 minute to study the picture series shown to them. After that, they proceeded to do the speaking test individually for approximately 4 minutes. At the end of the test, the test's assessor would not reveal their scores. All of their performance were recorded and sent to the juries. They used the same agreed rubric to score the test takers. They have been trained several times before finally come to judge the test-takers.

The Test's Reliability Results

To find the test's reliability, the writers needed to get the numerical data from this study to count the test-retest and interrater values. Therefore, during the test's tryout with the trial group, the writers asked three raters to score the try out test-takers performance. Test-retest reliability is a way to prove that measurement is stable and consistent enough to be used. It was proved by delivering the same test to the same test takers twice on separate occasions and times (Dutil, Bottari, & Auger, 2017). The interrater reliability, on the other hand, dealt with what extent an agreement among data collectors (May, 2006). In many types of research, multiple people were collecting and interpreting data together. This fact may lead to a biased result if it was not maintained carefully. By finding the interrater reliability, researchers may notice the agreement between raters used to make sense of the results at the end. The tables below provided information about the test-takers' scores on their first and final trial.

Table 1.

First trial results

FIRST TRIAL										
	First picture			Second picture			Average Scores			Average Scores
Students's Name	The writer	Rater A	Rater B	The writer	Rater A	Rater B	The writer	Rater A	Rater B	
Student 1	88	88	87	88	88	88	88	88	87	88
Student 2	90	89	89	90	89	89	90	89	89	89
Student 3	93	92	93	91	91	91	92	91	92	92
Student 4	94	94	93	93	93	93	94	93	93	93
Student 5	93	90	91	87	88	88	90	89	89	89
Student 6	88	87	87	91	89	90	89	88	89	89
Student 7	95	95	95	95	95	94	95	95	94	95

Table 2.
Final trial results

FINAL TRIAL										
	First picture			Second picture			Average Scores			Average Scores
Students's Name	The writer	Rater A	Rater B	The writer	Rater A	Rater B	The writer	Rater A	Rater B	
Student 1	90	90	88	88	89	89	89	89	88	89
Student 2	91	93	93	88	91	91	89	92	92	91
Student 3	93	93	92	91	91	91	92	92	92	92
Student 4	91	91	89	89	92	92	90	91	90	90
Student 5	86	85	89	87	89	90	86	87	89	87
Student 6	91	90	92	91	91	92	91	91	92	91
Student 7	95	95	94	95	94	94	95	95	94	94

Test-retest Reliability

The test's test-retest reliability was the first thing that the writers analysed. First of all, they made a table that can compare test-takers' first and final scores. After that, they counted each student's average scores from all raters. Then, they counted the test-retest value using Pearson's correlation formula. The writers did this step by themselves after doing enough research on how to do it properly using Excel. After checking the results, they proceeded to interpret the value found. The table presenting the result as followed.

Table 3.
Test-retest reliability results

TEST-RETEST		
Students' Name	Test 1	Test 2
Student 1	88	89
Student 2	89	91
Student 3	92	92
Student 4	93	90
Student 5	89	87
Student 6	89	91
Student 7	95	94
CORRELATION		0.7115794203

Table 4.
Test-retest reliability criteria

Less than 0.20	Slight, almost no relationship
0.21-0.40	Low, correlation; definite but small relationship
0.41-0.70	Moderate correlation; substantial relationship
0.71-0.90	High correlation; strong relationship
0.91-1.00	Very High correlation; very dependable relationship

The writer used a range of criteria to interpret the result above provided by (Guilford, 1956). It was shown as followed.

The result of the test's test-retest reliability showed that it has a high correlation relationship with the value of 0.71. It has been proved by this data that the test developed was highly reliable to be used. This result provided an answer to the first research question, which was to what extend is the test reliable.

Interrater Reliability. Interrater reliability was the second thing to analyse. The writers collected the scores taken from 3 raters. They put them on the same table. There were two scores typed in the table. They were the students' first and second test average scores. After compiling those scores, they then counted each rater's average scores for each test taker. They used the formula of average on Excel to count this. The table presented the scores and calculations as followed.

Table 5.
Scores Taken From Different Raters

Students' Name	Test 1			Test 2			Average Score		
	The writer	Rater A	Rater B	The writer	Rater A	Rater B	The writer	Rater A	Rater B
Student 1	88	88	87	89	89	88	88	88	88
Student 2	90	89	89	89	92	92	90	90	90
Student 3	92	91	92	92	92	92	92	92	92
Student 4	94	93	93	90	91	90	92	92	92
Student 5	90	89	89	86	87	89	88	88	89
Student 6	89	88	89	91	91	92	90	89	90
Student 7	95	95	94	95	95	94	95	95	94

After getting the average scores from each rater for each test taker, the writer made another table to count the interrater reliability. First of all, she made a table consisted of three rows namely students' names, average score, and difference pair. The students' names and average score rows on table 6 were taken from table 5. On the difference pair row, the writers needed to pair her raters before counting any further. Since this study used three raters, it has three different pairs in total. The first pair was the writer and rater A, the second pair was the writer and rater B and the last one was rater A and rater B. After that, the writers needed to find the score gaps between each pair. It was counted by doing a subtraction. In doing this, they did not do it manually. They used a formula in her excel to help them counting. After getting the score gaps, they needed to count the 0 values found there. The 0 value represented the raters' agreement; there were no score gaps found in the scoring. The presentation table as followed.

Table 6.
Interrater Reliability

Students' Name	Average Score			Difference pair		
	The writer	Rater A	Rater B	Writer and A	Writer and B	A and B
Student 1	88	88	88	0	0	0
Student 2	90	90	90	0	0	0
Student 3	92	92	92	0	0	0
Student 4	92	92	92	0	0	0
Student 5	88	88	89	0	-1	-1
Student 6	90	89	90	1	0	-1
Student 7	95	95	94	0	1	1
Total count of 0 in difference column				6	5	4
Total Rating				7	7	7
Proportion of Agreement				0.8571429	0.7142857	0.5714286

The result of the score gaps counting was presented in table 6. First of all, the writers needed to count the 0 values among the pairs. The writer and rater A have 6 zeros, the writer and rater B have 5 zeros and rater A and B have 4 zeros. After that, she counted the proportion of agreement from each pair. She divided the total count of zero with the total ratings of this study. The total rating of this study was 7. It was taken from the total test-takers who participated in the trial. The result of the calculation was put in the proportion of agreement column. As shown in Table 6, the writer and Rater A got a proportion of agreement of 0.85, categorised as an almost perfect agreement. The second

pair, the writer and Rater B, got a proportion of agreement of 0.71, categorised as a substantial agreement. The last pair, Rater A and B got a proportion of agreement of 0.57, categorised as a moderate agreement. The writer interpreted the data found using criteria provided by Landis and Koch (1977) as followed.

Table 7.
Interrater Reliability Criteria

< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

The test's Validity, Practicality and Authenticity Results

The test's validity, practicality, and authenticity were proved by providing an expert judgement about the test. The expert judgement used in this study was the Cambridge curriculum head of the school participated in this study. They were collected through questionnaires filling. The following were the aspects used in the questionnaires.

Table 8.
Aspects in the questionnaire

Validity	Practicality	Authenticity
Face Validity	Time	Realistic
Content Validity	Media	Universal
Construct Validity	Test Procedure	Transparent
	Effort	Engaged
	Result	Trustworthy

The overall judgement about the test was satisfying. It can be concluded that the writer has made a valid, practical, and authentic test. In the validity questionnaire, it can be concluded that the test made has great a great face, content, and construct validity. There, the expert has strongly agreed to most of the statements. First of all, in face validity, she was strongly agreed that the test was suitable, doable, well-developed, and interesting. In the construct validity, she also was agreed that the test's task matched the skill required, objective set, and curriculum used. In the content validity, she was agreed that the test task covered the recent materials, chapter, and topic discussed by targeted test takers. In the practicality questionnaire, she was strongly agreed that the time, media, test procedure, effort, and a result of the test were very efficient and suitable to be done in the recent pandemic learning situation. In the authenticity questionnaire, she was strongly

agreed with all of the elements used in it. It covered the test's realism, universality, transparency, engagement, and trustworthiness. The test developed was appropriate to the targeted test takers. It asked the test-takers, to be able to perform a universal skill that can be used outside the class. It also has a transparent scoring standard that test-takers can check before and after their performance. Moreover, they can access the scoring criteria, system, and rubric used to score them before their performance to prepare themselves better. All in all, the expert has agreed that the test developed has passed all of the requirements to be called a valid, practical, and authentic test.

Conclusions

From all the results above, it can be concluded that the test developed was reliable, valid, practical, and authentic. It can be used to targeted test-takers that share the same curriculum, level materials, and topics. It was reliable because the value of its test-retest and interrater reliability was high. The test-retest reliability was 0.71. It was categorised as a highly reliable test. As the test's interrater reliability was at the value of 0.85, 0.71, and 0.57. It can be categorised as an almost perfect, substantial, and moderate agreement. For the validity, practicality and authenticity were proved by the expert judgement through a questionnaire result. It was concluded that the test developed has passed her standard and worth to be used.

References

Aboagye, E., Yawson, J. A., & Appiah, K. N. (2021). COVID-19 and E-learning: The challenges of students in tertiary institute-ions. Social Education Research, 1-8.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

COVID-19 and higher education: Today and tomorrow. (2020, April 9). IESALC. http://www.iesalc.unesco.org/en/wp-content/uploads/2020/04/COVID-19-EN-090420-2.pdf

COVID-19 Assessing the impact on the education sector and looking ahead. (2020, April). EY PARTHENON. https://www.ey.com/Publication/vwLUAssets/ey-COVID-19-Assessing-the-impact-on-the-education-sector-and-looking-ahead/$File/ey-COVID-19-Assessing-the-impact-on-the-education-sector-and-looking-ahead.pdf

Diana, L. (2021). Problems Faced in Speaking Assesment During The Covid-19 Pandemic. Study Case Of Universitas Pembangunan Nasional Veteran Jawa Timur. Jisip (Jurnal Ilmu Sosial dan Pendidikan), 5(1).

Djafar, R. (2020). Analysis of The Effect of Covid-19 Towards L2 English Speaking Performance. JISIP (Jurnal Ilmu Sosial dan Pendidikan), 4(4).

Dutil, É., Bottari, C., & Auger, C. (2017). Test-retest reliability of a measure of independence in everyday activities: The ADL profile. Occupational therapy international, 2017.

Guangul, F. M., Suhail, A. H., Khalit, M. I., & Khidhir, B. A. (2020). Challenges of remote assessment in higher education in the context of COVID-19: a case study of Middle East College. Educational Assess-ment, Evaluation and Accountability, 1-17.

Guilford, J. P. (1956). The structure of intellect. Psychological bulletin, 53(4), 267.

Hasan, N., & Khan, N. H. (2020). Online Teaching-Learning During Covid-19

Pandemic: Students'perspective. The Online Journal of Distance Education and e-Learning, 8(4), 202.

Haynie III, W. J. (2003). Effects of take-home tests and study questions on retention learning in technology education. Volume 14 Issue 2 (spring 2003).

Huebner, Theodore. (1960). Audio Visual Technique in Foreign Language. New York: Cambridge University Press.

Krishan, I. A., Ching, H. S., Ramalingam, S., Maruthai, E., Kandasamy, P., De Mello, G., ... & Ling, W. W. (2020). Challenges of Learning English in 21st Century: Online vs. Traditional During Covid-19. Malaysian Journal of Social Sciences and Humanities (MJSSH),5(9), 1-15.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics, 363-374.

Larson, J. W. (1984). Testing Speaking Ability in the Classroom: the Semi-direct Alternative. *ForeignLanguage Annals*, 17(5), 499-507.

Lee, H.-W. (2020). Technology-Enhanced Language Assessment: Innovative Approaches for Better Learning. Cambridge Assessment https://www.cambridgeassessment.org.uk/insights/technology-enhanced-language-assessment-innovative-approaches-for-better-learning/

May, L. A. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. Melbourne Papers in Language Testing, 11(1), 29–51.

Nunan, D. (1999). *Second Language Teaching & Learning*. Heinle & Heinle Publishers, 7625 Empire Dr., Florence, KY 41042-2978.

Setyawan, C., & Aryati Prasetyarini, M. P. (2020). Challenges On Teaching Online English Subject In SMK Negeri-1 Nawangan (Doctoral dissertation, Universitas Muhammadiyah Surakarta).

Wang, Y. (2004). Supporting synchronous distance language learning with desktop Video conferencing. Language Learning & Technology, 8(3), 90-121.

Pictures used in the test: First: https://www.pinterest.com.au/pin/422071796329377577/ Second: https://www.pinterest.com.au/pin/422071796329377582/

About the Author

Yuditha Putri Wiwaharini has got her Master's degree from the English Education Department, Graduate School, Widya Mandala Surabaya Catholic University. She is a curriculum vice-principal at Nation Star Academy Surabaya. Her research focuses on English teaching and learning to gain insight to improve the students' interest in learning it. Recently, her work focused on developing adaptative media to teach English.