JERE 9 (1) (2020) 52-57

Journal of Educational Research and Evaluation

http://journal.unnes.ac.id/sju/index.php/jere

Content Validity and Reliability of The Inter-Rater Instrument for Android-Based Speaking Performance Assessment

Febri Dhany Triwibowo ✉, Ani Rusilowati, Dwi Anggani Linggar Bharati

Universitas Negeri Semarang, Indonesia

Info article

Published 31 March 2020

Keywords:
Content validity, content reliability, Speaking Performance Assessment Instrument

Abstract

This research is part of development research based on problems encountered in the field, which is the absence of an android-based speaking assessment instrument and one that supports one of the conservation values of Semarang State University, which is paperless. The aim of this study was to show the results of the validity and reliability test of the developed android-based speaking performance assessment instrument. The research method used is quantitative description of 3 expert judgments. The validation instrument developed in the form of an expert assessment sheet with 4 criteria, there are ease of use, visuals, content, and benefits. Analysis of the content validity of the assessment sheet using the V coefficient by Aiken and the reliability of the instrument content using the Interclass Correlation Coefficient (ICC) analysis with the help of SPSS version 16.0. The results showed valid results with all criteria valued> 0.3, namely with the lowest index 0.8 and the highest 1.0. Inter-rater reliability test using ICC obtained a value of 0.875, which means that all the criteria for the developed android-based speaking performance assessment instrument have a good level of consistency. Based on the results, it can be showed that the android-based speaking performance assessment instrument can be used.

✉Correspondence address :
Kelud Utara 3 kampus Pascasarjana UNNES Semarang, Indonesia
E-mail : febri.dhany96@gmail.com

p-ISSN 2252-6420
e-ISSN 2503-1732

INTRODUCTION

Tests and assessments have the same objective, which is to assess students' abilities but in different implementation. Tests and assessments have clear differences, such as tests are scheduled activities written in the school curriculum, and aim to determine the peak ability of students (Brown, 2004: 4). In every assessment of language activities, the clarity of the assessment indicators is very important to pay attention to before the test is made (O'Sullivan, 2013: 156). To know the abilities and attitudes of students as a whole, it is necessary to have a performance appraisal, which in this study focuses on speaking performance (Rusilowati, 2014: 781). This is in line with Masrukan's (2017: 6) opinion that assessment is an activity of systematically collecting data and analysing the information obtained for later use as material for improving decision-making.

One of the learning activities that requires non-test assessment is speaking. English as an international language is needed to support easier global communication. The ability to speak has an important role in society because it is a way of communicating, expressing ideas, opinions, messages, feelings, thoughts, and socializing with others (Putra, 2017: 36). On a global scale, speaking ability is increasingly in demand, and is sometimes recognized as a highly sought-after ability and as a benchmark for ability in the current educational context and work environment (Isaacs, 2016: 131). Because it is an important ability, in the context of learning, assessment of speaking skills must be taken seriously.

The problem is in the assessment of speaking performance which is considered as one of the types of assessments that are difficult to do in learning second and foreign languages (Beltrán, 2016: 1). Speaking ability is an ability that is difficult to assess objectively and its reliability because it involves a combination of different variables that have little or no correlation with each other (Ulker, 2017: 135). Most of the complexity of the speaking assessment comes from the different point of view of the construct definition, about what is required in the speaking assessment and how to assess it (Fulcher, 2013: 322).

The research developed was an android based assessment rubric. Like other assessment instruments, rubrics have different purposes and procedures for use with others, the main purpose of scoring using a rubric is to assess performance (Brookhart, 2013: 4). In designing the speaking assessment test, it must be based on an assessed activity model so that the type of assessment chosen will be assessed based on a ranking scale that is in line with the expected linguistic behaviour (Beltrán, 2016: 2). The rubric that was developed was an analytic type. Analytical or descriptive rubrics have separate criteria, have a scale and a description of each assessment component (Berger, 2011: 174).

The results of a preliminary study through interviews that conducted at the Faculty of Language and Arts, Semarang State University to several students who are taking or have taken the speaking course and the teaching lecturers of the speaking course, it can be concluded that there is no speaking assessment instrument based on the android application. Moreover, some students said that the assessment carried out in speaking activities was not transparent, which means that students did not know what aspects were assessed and how many grades were obtained. This strengthens researchers to develop an android-based speaking performance assessment instrument. The application developed will support the conservation value of Semarang State University, which is paperless and support lecturers to show the value of students and their indicators easily through the results of screenshots.

In accordance with the description above, an android-based speaking performance assessment instrument needs to be developed to assist in assessing speaking activities. Before being used, the speaking performance assessment instrument made in the form of an Android application must be analysed first to determine its quality. The formula used in the content validity test is the Aiken's V validity index because the instrument was tested or assessed by 3 experts. As for the interrater reliability test using the Interclass Correlation Coefficient (ICC) analysis.

METHODS

The method used in this research is descriptive quantitative because the data obtained from the results of the content validity and reliability tests. The data on the validity instrument content was obtained by providing an assessment questionnaire to 3 experts with different backgrounds, such as 1 research and evaluation expert, 1 English linguist, and 1 technology and informatics expert. There are 4 criteria assessed by experts, the criteria for user ease, appearance or visuals, content, and benefits. The results of the three experts' assessments were then analysed using the Interclass Correlation Coefficient (ICC) analysis. Based on these calculations, the value of content validity and interrater reliability of the problem-solving ability assessment instrument can be obtained.

Content validation is the validity that is estimated through testing the test content with rational analysis or through professional judgment (Rusilowati, 2014: 23). (Aiken, 1985: 133) formulates a formula for calculating the content-validity coefficient which is based on the results of an expert's assessment of n people on an item regarding the extent to which the existing items can represent the construct being measured or commonly referred to as Aiken's V. The assessment is carried out by giving a number1 to 5, the number 1 to represent the assessment is irrelevant until 5 represents the most relevant (Azwar, 2015: 112-113). The content validity test used the Aiken's V formula. The Aiken's V formula used was as follows:

$$V = \frac{\Sigma \mathrm{s}}{\mathrm{n}\ (c-1)}$$

Notes:

S = r - lo

Lo = the lowest number of validity assessments (= 1)

C = the highest number of validity assessments (= 4)

R = number given by an assessor

N = number of evaluators

If the validity coefficient is less than 0.30, it means that the item is inadequate (invalid), on the other hands, if the validity coefficient is ≥ 0.3, it means that the item is adequate or valid (Azwar. 2015: 134).

The reliability of the instrument content developed using the ICC approach with the help of SPSS 16.0 to calculate the level of agreement between the five Expert Judgments, before using the ICC approach, the reliability value was estimated using the Cronbach Alpha coefficient. The coefficient value must be more than r_{xx} > 0,5 because ICC less than 0.5, it indicates low reliability, while values between 0.5 and 0.75 indicate fairly good reliability, a value between 0.75 and 0.9 indicates good reliability, and a value of more than 0.90 indicates perfect reliability. (Portney & Watkins, 2009: 186).

RESULTS AND DISCUSSION

The validity assessment was carried out by experts with different backgrounds. There are 1 research and evaluation expert, 1 English linguist, and 1 informatics expert. The expert appointed makes an assessment of two main things, there are assessing whether the grid is made showing that the classification of the grid represents the aspects and indicators to be assessed, namely speaking performance and after that the expert assesses whether each of the criteria in the speaking performance assessment application is relevant with the classification grid that has been determined and in accordance with the product being made. Analysis of the content validity of the instrument used the Aiken coefficient V, where V index> 0.3. The following is Table 1 of the results of the content validity test using the Aiken V formula.

Table 1. Content Validity Results using Aiken V

No. Items	Rater 1	Rater 2	Rater 3	V indeks	Remarks
1	4	4	4	1,0	Valid
2	3	4	3	0,8	Valid
3	4	4	4	1,0	Valid
4	4	3	4	0,9	Valid

(Source: Researcher Data, 2020)

In Table 1, the results of the Aiken V analysis produce all items that are declared valid with the lowest index 0.8 and the highest index of 1.0. These results indicate that all items are declared valid.

Through the validation questionnaire also shows qualitative data from the validator in the form of suggestions and input which are notes for researchers to make improvements to the instruments and products developed. From the expert suggestions, revisions are made so that the application product can have a more attractive appearance, improve grammar on the instrument, and provide the option to save the assessment results document. The advice given by experts on the instrument can be seen in Table 2.

Table 2. Assessment Results and Suggestions from the Validator

Experts Name	Comments/Suggestion
Expert 1	Revised grammar on the *instruction*
Expert 2	*Instruction* made with interesting impressions (photos / moving images)
Expert 3	Give initial info a description of the application and what can be done to get started Can it not be documented / stored?

(Source: Researcher Data, 2020)

The suggestion that provided by the expert will later be used as a reference in improving instruments and applications before small and large scale trials are carried out. Improvements include improving grammar in the application, especially in the instruction section, besides checking grammar in other parts of the application. In the instruction section there are also suggestions for adding images, it can make the interface will be more attractive. The last expert's suggestion is to provide a value storage feature after the user will conduct an assessment in class.

The reliability test between validators was tested using the ICC (Interclass Correlation Coefficient) formula using SPSS 16.0 to estimate the interrater reliability by showing the comparison between variations caused by the attributes measured with the overall measurement variation. The calculation of the content reliability using SPSS 16 can be seen in Table 3.

Table 3. Validity Test Result by 5 Expert

Cronbach's Alpha	N of Items
.875	3

(Source: Researcher Data, 2020)

From the results of the reliability test above, it can be seen that the range of reliability of the assessment instrument developed is> 0.5, which means that it has met the good reliability requirements. According to Khumaedi (2012: 29) that the reliability coefficient> 0.5 is quite accepted as good reliability. Based on Cronbach's Alpha analysis, it shows that the result of agreement interrater is the coefficient value r_{xx} = 0,874, so the coefficient value r_{xx} > 0,7 it can be continued with the ICC (Interclass Correlation Coefficient) analysis with the help of SPSS 16.The following are the results of the ICC test for agreement interrater in Table 4.

Table 4. Interclass Corellation Coeficient Agreement of Inter-Rater

Interclass Correlation Coefficient	
	Intraclass Correlation[a]
Single Measures	.700[b]
Average Measures	.875[c]

Based on the ICC output results above from Single Measures = 0.700. Based on the Portney & Watkins classification, the ICC value between the ranges of 0.5 to 0.75 indicates a fairly good reliability (2009: 186). Therefore the assessment instrument developed and tested for validity and reliability by 3 experts using the Aiken V formula and reliability using Interclass Correlation Coefficient (ICC) analysis produced valid and reliable results.

Through the results of the validity and reliability tests that have been carried out, it can be said that the application of the speaking performance assessment instrument developed is valid and reliable for use. This study produced an assessment instrument that supports the conservation value of the State University of Semarang (UNNES), namely reducing paper use in lecture activities and administration. Previously, there was no speaking performance assessment application developed in the scope of learning English at UNNES so this research provides new objectives for the development of speaking performance assessment instruments in the future. In addition, by developing this application, researchers hope that this application can later be used and helps make it easier for lecturers to make paper-free assessments.

CONCLUSION

The speaking performance assessment instrument and the developed android application product were declared feasible after going through the content validity test by the expert and analysed with the Aiken V formula with the lowest index result of 0.8 and the highest index of 1.0. This shows that the aspects and criteria of the speaking performance assessment instrument developed are valid.

Through the reliability test, the speaking performance appraisal instrument also had good reliability after getting a value of 0.700 with the Interclass Correlation Coefficient (ICC) analysis assisted by SPSS 16.0. Thus the speaking performance assessment instrument and the developed android application have a fairly high agreement Interrater.

ACKNOWLEDGMENT

The author would like to thank Girindra Putri Dewi Saraswati, S.Pd., M.A., as expert validator 1, Prof. Dr. Totok Sumaryanto Florentinus, M.Pd. as expert validator 2, and Dr. Budi Naini Mindyarto, M. App. Sc., As expert validator 3 who has provided an assessment of the android-based speaking performance assessment instrument that has been developed.

REFERENCES

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurment*, *5*(1), 131–142.

Azwar, S. (2015). *Reliabilitas dan Validitas*. Pustaka Pelajar.

Beltrán, J. (2016). The Effects of Visual Input on Scoring a Speaking Achievment Test. *Studies in Applied Linguistics and TESOL*, *16*, 1–2.

Berger, J. (2011). Evaluating the Effectiveness of Instruction Using Principles of Adult Learning. In *Assessing and Evaluating Adult Learning in Career and Technical Education* (pp. 173–190). Idea Group Inc.

Brookhart, S. M. (2013). *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD.

Brown, H. D. (2014). *Language Assessment Principles and Classroom Practices* (2nd ed.). Pearson Education, Inc.

Budhiwaluyo, N., Asyhar, R., & Hariyadi, B. (2016). Pengembangan Instrumen Penilaian Kinerja pada Praktikum Struktur dan Fungsi Sel Di SMA Negeri 1 Kota Jambi. *Edu-Sains: Jurnal Pendidikan Matematika Dan Ilmu Pengetahuan Alam Universitas Jember*, *5*(2).

Fulcher, G., & Reiter, R. M. (2013). *Task Difficulty in Speaking Test*. *20*(3), 321–344.

Isaacs, T. (2016). Assessing Speaking. *Handbook of Second Language Assessment*, *12*, 131–146.

Khumaedi, M. (2012). *Reliabilitas Instrumen Penelitian Pendidikan*. *12*(1), 29.

O'Sullivan, B. (2013). *Assessing Speaking*. *1*, 156–171.

Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: Applications to practice* (Vol. 892). Pearson/Prentice Hall.

Putra, A. (2017). The Correlation Between Motivation And Speaking Ability. *Channing: Journal of English Language Education and Literature*, *2*(1), 36–57.Rusilowati, A. (2014). *Pengembangan Instrumen Penilaian*. Semarang Press.

Ulker, V. (2017). The Design and Use of Speaking Assessment Rubrics. *Journal of Education and Practice*, *8*(32), 135–141.