IJAZ ARABI: Journal of Arabic Learning
DOI: 10.18860 /ijazarabi.v5i1.13854
ISSN(print): 2620-5912 | ISSN(online): 2620-5947
ejournal.uin-malang.ac.id/index.php/ijazarabi/index | 210

Reliability Assessment of Arabic Speech Contest

Andhita Dessy Wulansari
Institut Agama Islam Negeri Ponorogo, Indonesia
andhita@iainponorogo.ac.id

Abstract

This study's assessment model for the Arabic speech competition involved three competent judges. Each of them assessed the same three components, namely *al fashahah* (fluency), *lubbul maudhu'* (content/theme discussion), and *al harakah* (participant movement, including participant expressions). The subjects in this study were all 24 participants in the Arabic speech competition at IAIN Ponorogo in 2019. All of these participants came from MA around Madiun. It is possible to issue differences between scores given by the three judges, considering that the range of values for each component is between 50 to 100. Scoring with a reasonably wide range could affect the consistency of the assessment. It is crucial to estimate the reliability coefficient of the judges' assessment of the Arabic speech contest. This study uses a quantitative approach because the primary data are scores from the judges of the Arabic language competition. The reliability estimation uses a variance analysis approach whose procedure is based on generalizability theory (G-Theory) through the G-Study concept with a multifaceted design. This theory can improve the instrument's quality by testing several sources of variance to make decisions and the consistency of the results of the generalizability coefficient. The data analysis results conclude that the assessment instrument used in the Arabic speech competition in this study is reliable or still feasible to use. The feasibility is based on the validity and reliability of the instruments used. The analysis of some experts regarding the instrument content claims that it was valid. In addition, the reliability coefficient of the combined score of the Arabic speech competition assessment is 0.96708. Therefore, it concludes that the assessment instrument used in this study is reliable.

Keywords: Reliability; Generalizability Theory; Speech Contest; Arabic

INTRODUCTION

The Arabic speech contest is inseparable from measurement and assessment as a competition. Measurement and assessment have different substances, but both are bound by a hierarchical relationship(Griffin & Nix, 1991, p. 3). Unfortunately, research related to the assessment of the Arabic language competition has rarely been done before. In 2021, research related to the Arabic language competition was an investigation done by Fikri et al. However, in this study, the focus is more on finding out what strategies are needed to increase the participants' confidence in the Arabic debate competition at the international level. According to this research, it is apparent that the debate competition contributes to increasing three competencies, namely linguistic, communicative, and cultural competencies more quickly and simultaneously, with two aspects, namely obtaining material and presenting speeches (Fikri, Machmudah, Halimi, & Ibrahim, 2021, p. 1).

Assessment is a crucial component in the delivery of education(Mardapi, 2012, p. 12). It is always carried out after doing a measurement. A measurement means an activity of assigning a score to an individual or a characteristic based on specific rules (Ebel & Frisbie, 1986, p. 14). It can also be interpreted as an activity to systematically set a score for an individual to state how he is doing (Allen & Yen, 2001, p. 2). On the other hand, assessment is interpreting the measurement data. It is the process of systematically collecting data to make decisions about a person (Berk, 1986, p. 16). Assessment includes all ways to collect data about an individual so that the decision is also on that individual (Mardapi, 2012, p. 13).

The assessment carried out was to decide the Arabic speech participants' quality. Score differences in the measurement results are very likely to occur in a competition, so the interpretation is also individual. Therefore, the assessment focuses on the individual. (Mardapi, 2008, p. 3) argues that assessment in the education field focuses on the learning outcomes achieved by each student. The crucial role of the assessment in learning outcomes is to determine learning outcomes (Suskie, 2018, p. 33). The way teachers assess has a considerable impact on learning. The way teachers assess has a considerable impact on learning. A proper assessment can positively impact the quality of learning outcomes. Eventually, proper assessment is considered a means of helping students to learn, reporting student progress, and making decisions about teaching (Jimaa, 2011, p. 721). The assessment in this study focuses on the quality of the participants' Arabic speech. The assessment is determined from the accumulated scores obtained by the participants from the judges.

The Ponorogo State Islamic Institute (*IAIN*) is one of the Islamic Colleges (*PTKIN*) organizing an Arabic speech competition for students from *Madrasah Aliyah* (*MA*) in the Madiun. The participants were assessed for their proficiency in speaking Arabic. Participants' skills were acquired from their various lessons and training (Nurbayan, Nurbayan, & Falah, 2020, p. 275). Learning a foreign language is not an easy thing. Learning Arabic is no exception, and until now, it is being studied more and more in Indonesia (Abdurochman, 2017, p. 1). Learning Arabic is difficult for Indonesians because of its differences in pronunciation, grammar, and vocabulary from their daily Indonesian language (Ramadani & Baroroh, 2020, p. 292).

The assessment of the Arabic speech competition conducted at IAIN Ponorogo involved three judges, each of which judged the same three components. There are 3 components assessed by the judges: *al fashahah* (fluency), *lubbul maudhu'* (content/theme discussion), and *al harakah* (participant movements, including participant expressions). The participant who gets the highest accumulated score from the judges is the participant who has the best Arabic speech quality and therefore he deserves to be the champion. It is possible to issue differences between scores given by the three judges, considering that the range of values for each component is between 50 to 100. Scoring with a fairly wide range could affect the consistency of the assessment between the judges. So that to improve the design of the assessment model, it is crucial to estimate the reliability coefficient of the assessment of the Arabic speech contest by the judges.

Reliability comes from 2 words, namely rely and ability. Reliability is related to the level of confidence, regularity, consistency, and stability of the measurement results of specific instruments (Saifudin Azwar, 2010, pp. 4-5). The core concept of reliability is how a measurement result can be trustworthy (Saifudin Azwar, 2010, pp. 4-5). Reliability can also be interpreted as a coefficient showing the consistency of measurement results (Mardapi, 2012, p. 52). The reliability of an instrument is related to the consistency of the measurement results for the same research variable while the measurement time is different. Evidence of reliability is ensures the consistency or constancy of the measurement results since a measurement result must be relatively the same if the measurement is applied to the same research variable, even though it is applied to a different respondent, time, and place. An instrument both test and non-test will be considered reliable if the measurement results are relatively the same for the same variable. Relative here does not mean that it has to be the same. There may be changes in the results, but they are not significant, and therefore, they are ignorable. Based on the data collection method, reliability is classified into three, namely: (i) internal consistency; (ii) stability; and (3) understanding between raters (Mardapi, 2012, p. 52). The reliability evidence of interrater consistency in this study was using the ANOVA (Analysis of Variance) table or it is more familiar with the name reliability with generalizability theory (Theory-G). The reliability value is referred to as the reliability coefficient. G-theory is a comprehensive procedure for designing, assessing, improving the internal consistency and stability of measurements (Williams, Roggenbuck, Patterson, & Watson, 1992).

Fisher first introduced the generalizability theory (G-theory) in 1925. It is an estimation concept with ANOVA which classifies observation conditions in various aspects. G-theory based on ANOVA can recognize various sources of error. It considers the effect of changing the condition of the measurement facets based on different multifaceted designs (Woodward & Joe, 1973, p. 173). The concept of G-Theory is a refinement of the concept of reliability with classical test theory. Reliability in classical test theory generalizes a sample into a randomly taken population of observations (Matt et al., 2000). Since the sample is generalized into many different observational populations, the estimation of the reliability coefficient is actually no longer accurate to the actual conditions. G-theory has several advantages over classical test theory in estimating reliability coefficients: (i) It can show a simultaneous overall measurement error; (ii) The estimation considers the effect of measurement error due to the interaction factor between components; (iii) Estimating the reliability coefficient by determining the ratio of the actual person variance to the observed person variance (Izza, Susilaningsih, & Harjito, 2014, p. 33). Besides identifying random errors, G-theory can also identify systematic errors so that it can present the dimension of validity and reliability together. It, of course, cannot be done with classical test theory, since in classical test validity and reliability are two different concepts and therefore the estimation is also done separately (A. D. Wulansari, Kumaidi, & Hadi, S., 2019, p. 141). In classical test theory, reliability is a test measure of random error, while validity is a test measure of systematic error.

GENOVA (A Generalized Analysis of Variance System) is a computer program application based on G-Theory developed by Robert L. Brennan in 1983. The G-theory here includes G study (generalized study) and D study (decision study). In its analysis, G-Theory is seen as a hierarchical process consisting of two levels. At the first level, the G study, component variance is estimated. The second level, the D study, employs the estimation, use, and interpretation of the estimated variance results to make decisions by using reliable measurement procedures. *D study* is used to obtain the reliability coefficient (Guntur, 2012, p. 155). According to Brannen, the D study emphasizes estimation, use, and interpretation to make decisions (Retnowati, 2012).

The researcher argues that the assessment of Arabic speech contests can be developed from the theory of measurement and assessment, as done by experts in the education field. Therefore, this study aims to discover the reliability of the Arabic speech contest assessment with G-Theory by following the measurement and assessment theory. The results of this study can be used as a recommendation to use the suitable assessment model for Arabic speech competitions in the future.

METHOD

The Arabic language competition involved three competent judges, each of whom assessed the same three components. The subjects in this study were all 24 participants in the Arabic speech competition at IAIN Ponorogo in 2019. All of these participants came from MA around Madiun.

The instrument used in this study is a scoring rubric. The scoring rubric was used to collect data on the participants' Arabic speech quality measured from three components, they are *al fashahah* (fluency), *lubbul maudhu'* (content/theme discussion), and *al harakah* (participant movements, including participant expressions). In addition, model testing includes estimates of validity and reliability. The validity evidence of the instrument was done by following the validators' judgment. The reliability estimation used a variance analysis approach whose procedure is based on generalizability theory through the G-Study concept with a multifaceted p x r x i design. It is because the variances of the competition assessment are based on three faceted variations; participants/person (p), judge/rater (r), and component/item (i). The reliability coefficient obtained was then compared with the minimum allowed reliability criteria, 0.7 (Linn, 1991, p. 143). The minimum reliability value is 0.7. If it is less than 0.7, the measurement error will exceed the limit (Basrowi, 2012).

This study employed a quantitative approach because the primary data are scores from the judges of the Arabic language competition at IAIN Ponorogo. The data analysis technique used descriptive analysis techniques. The descriptive data analysis technique was used to explain the characteristics of the Arabic speech competition assessment model at IAIN Ponorogo.

RESULTS AND DISCUSSION

The instrument used in this study is a scoring rubric to assess the participants' Arabic speech performance measured from three components. They are *al fashahah* (fluency), *lubbul maudhu'* (content/theme discussion), and *al harakah* (participant movement, including participant expressions). A performance is only suitable to be measured by a performance test, and Arabic speech is a performance, therefore the assessment of Arabic speech should use a performance assessment. Performance assessment is needed to assess the participants' skills and creativity (Stiggins, 1994, p. 171). The characteristics of the assessment were considered suitable to be used in the Arabic speech competition at IAIN Ponorogo. In assessing performance, the teacher wants the students' authentic responses of observable activities. In the Arabic speech competition, the measurement is carried out directly on the participants' performance in the Arabic speech so that the measurement results describe the contestants' actual abilities/skills.

The object of performance assessment is everything related to observable performance. Observable performance in the Arabic speech competition assessment is different from the learning assessment. In learning, observed performance relates to complex cognitive processes, for example working together, conducting experiments, measuring, analyzing, making decisions, and demonstrating a product. In addition, performance assessments can also be used to access students' ways of thinking, working, and behaviour in real life. This type of assessment follows learning effectiveness (Baker, 1997, p. 248). In the Arabic speech competition assessment, the observable performance includes all aspects assessed by the participants. All objects assessed in the Arabic speech contest are observable through hearing, feeling, and seeing.

Observation through hearing is carried out by assessing the *al fashah* (fluency) aspect. Observations through feelings are carried out by assessing the *lubbul maudhu'* (content/theme discussion) aspect. Observations through sight/seeing are carried out by assessing the *al harakah* (participants' movements, including expressions). In practice, the three types of observations are carried out by three competent judges simultaneously.

The instrument in conducting the Arabic speech contest assessment is not much different from the learning assessment. As the learning assessment, the performance assessment also uses an observation rubric. The observation rubric lists the aspects observed. Based on the descriptors seen during the observation process, the participants' performance scores in the Arabic speech contest were determined based on the pre-determined assessment criteria.

The Outline Of Assessment Instrument

Assessment of Arabic speech performance can be measured from three areas. They are *al fashahah* (fluency), *lubbul maudhu'* (content/theme discussion), and *al harakah* (participant movements, including participant expressions). These three areas are the components of the assessment in this study. The assessment is carried out through some indicators that are the scope of each.

Table 1. The Arabic Speech Contest Assessment Instrument Outline

Assessment Aspects	Indicators	Assessment Items
1. *al Fashahah*	a. Reading Accuracy	▪ The accuracy of reading the arguments (*dalil*) according to the law
	b. Translation Accuracy	The accuracy of translating the argument (*dalil*)
	The suitability of the arguments (*dalil*) with the topic	▪ The accuracy of the selected argument (*dalil*) with the specified topic
2. *Lubbul maudhu'*	a. Description Coverage	Ability to deliver a description under the scope of the problem specified Ability to convey descriptions systematically
	b. Description Systematics	▪ Ability to convey language utterance appropriately
	c. Language utterance d. Language style	Ability to convey language style appropriately
3. *al Harakah*	Vowel, Intonation, and Accent	Ability to produce sound properly and correctly The ability to present the high and low notes in sentences correctly Ability to put stress on syllables or words correctly
		▪ Ability to express feelings, intentions, or ideas appropriately
	b. Expression	▪ Ability to situate himself
	c. Manner	

Assessment cannot be separated from scoring activities, as it is in the Arabic speech competition assessment at IAIN Ponorogo. Scoring in this study was carried out for each indicator. The scoring was done to show each participant's Arabic speaking skills. Participants who get a high score mean that their Arabic speech ability is good, and vice versa, those who get a small score indicate that their Arabic speech ability is not good. Stating Arabic speech ability with a score will be easier to interpret than a narrative/qualitative assessment only.

Instrument Reliability

The instrument's feasibility is obtained by proving its validity and reliability (A. D. Wulansari, Kumaidi, Hadi, S., Saleh, M., & Friyatmi, 2019, p. 566). If the results of the verification show a coefficient value that is greater than or equal to the specified criteria, the instrument is considered valid and reliable (Fauziana & Wulansari, 2021, p. 17). It is, therefore, feasible to use. If the opposite happens, however, the results of the validity and reliability will generate a coefficient value that is smaller than the specified criteria. It is, therefore, considered invalid and unreliable. Validity is a measure of how accurately an instrument performs its measuring function (Mardapi, 2004, p. 25).

Proving the validity of the instrument used here is done through expert judgment. Proof of validity through expert judgment is carried out through rational analysis of the contents of a test instrument based on an individual subjective opinion of the experts (Allen & Yen, 2001, p. 95). Conducting such validity proof can also be referred to as professional judgment (Saifuddin Azwar, 2012, p. 45). Proving the instrument validity with expert judgment was applied because determining indicators, and assessment items in the Arabic speech competition at IAIN Ponorogo require the opinion of the experts in their fields. Judgment on the instruments' contents used in this study is carried out by Arabic language lecturers, as experts, at IAIN Ponorogo.

In contrast to the validity proof that has been done, the assessment of the Arabic speech contest reliability proof was assited by the GENOVA or SPSS program. The reliability estimation carried out in this study used an internal consistency approach. This approach is the most practical to apply because it uses a single test by applying the Analysis of Variance (ANOVA) and not the split-half method. Estimates of reliability using analysis of variance is logical because the concept of reliability itself is the ratio of various distribution variances (Saifuddin Azwar, 2012, p. 92).

The ANOVA procedure in this study is based on generalizability theory through the G-Study concept with a multifaceted p x r x i design because the variances of the competition assessment are based on three faceted variations. They are participant/person (p), judge/rater (r), and component/item. (i). These three faceted variations can produce seven different variance components. The seven components of the variance are (Thorndike, 1982, p. 161) :

σ^2_{pir} = person variance x item x rater

σ^2_{pi} = person variance x item

σ^2_{pr} = person variance x rater

σ^2_{ir} = item variance x rater

σ^2_{p} = person variance

σ^2_{i} = item variance

σ^2_{r} = rater variance

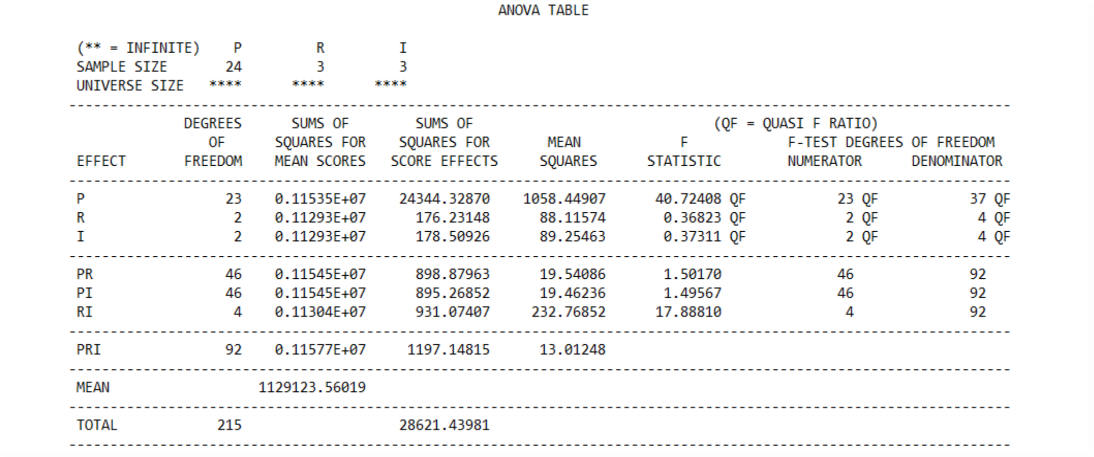
The procedure for estimating the reliability of the Arabic language competition assessment instrument at IAIN Ponorogo was done by finding the value of the variances above.

Table 2. The Mean Square of Competition Assessment Model Combined Score with SPSS

Source		Type III Sum of Squares	Df	Mean Square	F	Sig.
Intercept	Hypothesis	1129123.560	1	1129123.560	1155.634	.000
	Error	13298.670	13,611	.977,060[a]		
PERSON	Hypothesis	24344.329	23	1058.449	40,724	.000
	Error	955,449	36,761	25,991[b]		
RATER	Hypothesis	178,509	2	89,255	.373	.709
	Error	1009.886	4,222	239.218[c]		
ITEM	Hypothesis	176,231	2	88,116	.368	.712
	Error	1010.876	4,224	239.297[d]		
PERSON * RATER	Hypothesis	895,269	46	19,462	1,496	.052
	Error	1197.148	92	13.012[e]		
PERSON * ITEM	Hypothesis	898,880	46	19,541	1,502	.050
	Error	1197.148	92	13.012[e]		
RATER * ITEM	Hypothesis	931,074	4	232,769	17,888	.000
	Error	1197.148	92	13.012[e]		
PERSON * RATER * ITEM	Hypothesis	1197.148	92	13,012	.	.
	Error	.000	0	.[f]		

Reliability estimates are done by finding the value of each of the variances above. The GENOVA and SPSS programs were used to get the calculation results of the seven mean squares. The following is the output presented in the ANOVA table generated using SPSS in Table 1 and GENOVA in Table 3.

Table 3. The Mean Square of Competition Assessment Model Combined Score with GENOVA



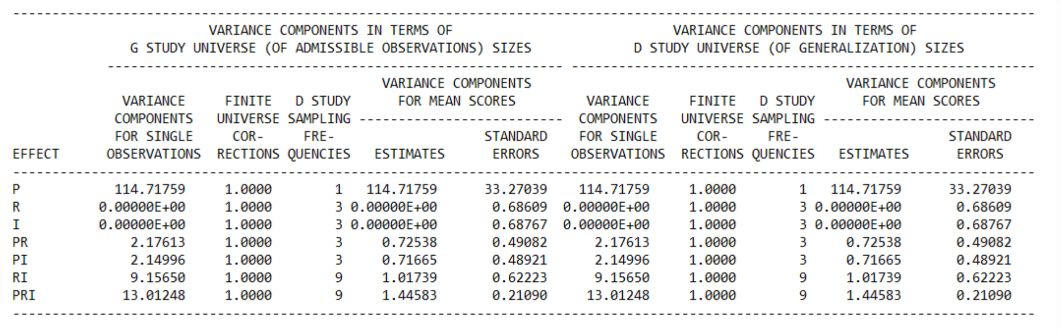
```
                                                     ANOVA TABLE

 (** = INFINITE)    P          R          I
 SAMPLE SIZE       24          3          3
 UNIVERSE SIZE   ****       ****       ****
---------------------------------------------------------------------------------------------------------------------
              DEGREES     SUMS OF        SUMS OF                            (QF = QUASI F RATIO)
                 OF      SQUARES FOR    SQUARES FOR      MEAN          F           F-TEST DEGREES OF FREEDOM
 EFFECT       FREEDOM    MEAN SCORES   SCORE EFFECTS    SQUARES    STATISTIC       NUMERATOR      DENOMINATOR
---------------------------------------------------------------------------------------------------------------------
 P                 23    0.11535E+07    24344.32870   1058.44907    40.72408 QF          23 QF           37 QF
 R                  2    0.11293E+07      176.23148     88.11574     0.36823 QF           2 QF            4 QF
 I                  2    0.11293E+07      178.50926     89.25463     0.37311 QF           2 QF            4 QF
---------------------------------------------------------------------------------------------------------------------
 PR                46    0.11545E+07      898.87963     19.54086     1.50170             46              92
 PI                46    0.11545E+07      895.26852     19.46236     1.49567             46              92
 RI                 4    0.11304E+07      931.07407    232.76852    17.88810              4              92
---------------------------------------------------------------------------------------------------------------------
 PRI               92    0.11577E+07     1197.14815     13.01248
---------------------------------------------------------------------------------------------------------------------
 MEAN                  1129123.56019
---------------------------------------------------------------------------------------------------------------------
 TOTAL            215                    28621.43981
---------------------------------------------------------------------------------------------------------------------
```

The mean squares values in Table 1 and Table 2 are proven to be the same therefore, it is possible to use them in calculating the value of the seven variance components mentioned.

When the mean square values in Table 1 and Table 2 are inserted into the seven variance formulas mentioned, the following is the result of calculating the values of the seven variance components with the help of the GENOVA program

Table 4. The Variance of Competition Assessment Model Combined Score with GENOVA



```
------------------------------------------------------------------------------------------------------------------------------
                       VARIANCE COMPONENTS IN TERMS OF                               VARIANCE COMPONENTS IN TERMS OF
             G STUDY UNIVERSE (OF ADMISSIBLE OBSERVATIONS) SIZES             D STUDY UNIVERSE (OF GENERALIZATION) SIZES
            -------------------------------------------------------- ---------------------------------------------------------
                                            VARIANCE COMPONENTS                                        VARIANCE COMPONENTS
              VARIANCE     FINITE   D STUDY   FOR MEAN SCORES        VARIANCE     FINITE   D STUDY       FOR MEAN SCORES
             COMPONENTS   UNIVERSE SAMPLING -------------------------  COMPONENTS   UNIVERSE SAMPLING -------------------------
             FOR SINGLE     COR-     FRE-                  STANDARD   FOR SINGLE     COR-     FRE-                  STANDARD
 EFFECT     OBSERVATIONS  RECTIONS QUENCIES   ESTIMATES     ERRORS   OBSERVATIONS  RECTIONS QUENCIES   ESTIMATES     ERRORS
------------------------------------------------------------------------------------------------------------------------------
 P             114.71759    1.0000        1   114.71759    33.27039     114.71759    1.0000        1   114.71759    33.27039
 R           0.00000E+00    1.0000        3 0.00000E+00     0.68609   0.00000E+00    1.0000        3 0.00000E+00     0.68609
 I           0.00000E+00    1.0000        3 0.00000E+00     0.68767   0.00000E+00    1.0000        3 0.00000E+00     0.68767
 PR              2.17613    1.0000        3     0.72538     0.49082       2.17613    1.0000        3     0.72538     0.49082
 PI              2.14996    1.0000        3     0.71665     0.48921       2.14996    1.0000        3     0.71665     0.48921
 RI              9.15650    1.0000        9     1.01739     0.62223       9.15650    1.0000        9     1.01739     0.62223
 PRI            13.01248    1.0000        9     1.44583     0.21090      13.01248    1.0000        9     1.44583     0.21090
------------------------------------------------------------------------------------------------------------------------------
```

Therefore:

$\sigma^2_{pir} = 13,01248$

$\sigma^2_{pi} = 2,149996$

$\sigma^2_{pr} = 2,17613$

$\sigma^2_{ir} = 9,15650$

$\sigma^2_{p} = 114,71759$

$\sigma^2_{i} = 0,00000$

$\sigma^2_{r} = 0,00000$

Reliability estimates using ANOVA are the ratio between various distribution variances. In this context, reliability estimates using ANOVA are the ratio or comparison between pure score variance and observed score variance (Thorndike, 1982, p. 163). The reliability coefficient in this study is the coefficient value obtained from the comparison between the pure score variance and the observed score variance of a test. Based on these equations, the reliability coefficient of the combined score can be defined as the value obtained from the comparison between the combined pure score variance and the combined observational score variance from the three areas of assessment for the Arabic speech contest, the *al fashahah* (fluency), *lubbul maudhu'* (content/discussion of the theme), and *al harakah* (participant movements, including participant expressions). Therefore, the reliability of the combined score (rxx') can be estimated using the following formula.

$$r_{xx'} = \frac{\sigma^2_{true}}{\sigma^2_{obs}}$$

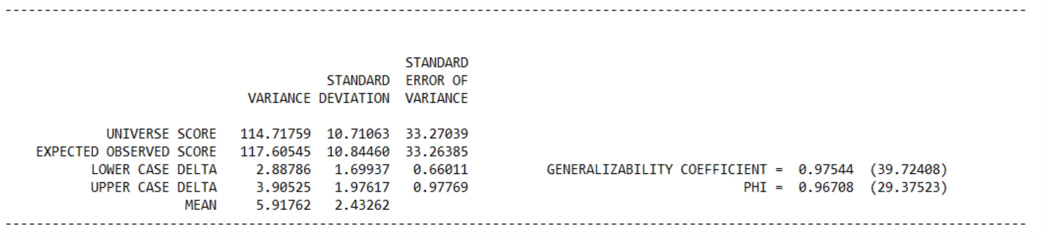
$$r_{xx'} = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_i}{n_i} + \frac{\sigma^2_r}{n_r} + \frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_{ir}}{n_i n_r} + \frac{\sigma^2_{pir}}{n_i n_r}}$$

$$r_{xx'} = \frac{114.71759}{114.71759 + \frac{0,00000}{3} + \frac{0,00000}{24} + \frac{2,149996}{3} + \frac{2,17613}{3} + \frac{9,15650}{9} + \frac{13,01248}{9}}$$

$$r_{xx'} = 0,96708$$

Then the coefficient of reliability of the combined score is the same as in the GENOVA output (Table 5).

Table 5. Reliability Coefficient of Combined Score using GENOVA



```
------------------------------------------------------------------------------------------------------------


                                                STANDARD
                                      STANDARD  ERROR OF
                            VARIANCE DEVIATION  VARIANCE

         UNIVERSE SCORE    114.71759  10.71063  33.27039
EXPECTED OBSERVED SCORE    117.60545  10.84460  33.26385
       LOWER CASE DELTA      2.88786   1.69937   0.66011        GENERALIZABILITY COEFFICIENT =  0.97544  (39.72408)
       UPPER CASE DELTA      3.90525   1.97617   0.97769                                 PHI =  0.96708  (29.37523)
                   MEAN      5.91762   2.43262
------------------------------------------------------------------------------------------------------------
```

From the comparison between the pure score variance and the data observation score, the reliability coefficient of the combined score of the Arabic speech competition assessment at IAIN Ponorogo was 0.96708. Because the value of rxx' is more than 0.7 (the minimum allowed reliability criterion), it concludes that the instrument used in the Arabic speech competition assessment at IAIN Ponorogo is reliable. Therefore, the assessment instrument used in the Arabic speech competition assessment at IAIN Ponorogo is feasible to be used.

CONCLUSION

In developing the assessment instrument for the Arabic speech competition at IAIN Ponorogo, it is very necessary to examine the theory of measurement and assessment as the basis for developing an assessment model. This study was carried out to discover whether the assessment model used is worth applying. One of the criteria is to see the reliability coefficient. Based on the results and discussion in this study, the reliability coefficient of the combined score for the Arabic speech competition assessment at IAIN Ponorogo was 0.96708. Therefore, the instrument used in the Arabic speech competition assessment at IAIN Ponorogo is reliable or means that the assessment instrument used in the Arabic speech competition assessment at IAIN Ponorogo is feasible to use. In conclusion, there is no need to make changes to the competition assessment instruments and techniques.

REFERENCES

Abdurochman, A. (2017). Strategi Pembelajaran Kosakata Bahasa Arab Bagi Non Arab. *19*(1), 63-83.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*: Waveland Press.

Azwar, S. (2010). Teori Sikap dan Pengukurannya. 155-157.

Azwar, S. (2012). Reliabilitas dan validitas.

Baker, E. L. (1997). Model-based performance assessment. *36*(4), 247-254.

Basrowi, S. (2012). Evaluasi Belajar Berbasis Kinerja. In: Bandung: Karya Putra Darwati.

Berk, R. A. (1986). *Performance assessment: Methods & applications*: Johns Hopkins University Press.

Briesch, A. M., Chafouleas, S. M., & Johnson, A. (2016). Use of generalizability theory within K–12 school-based assessment: A critical review and analysis of the empirical literature. *29*(2), 83-107.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *52*(1), 13-35.

Ebel, R., & Frisbie, D. (1986). Essentials of educational measurement. Englewood Cliffs, NJ: Prenctice-Hall. In: Inc.

Fauziana, A., & Wulansari, A. D. (2021). Analisis Kualitas Butir Soal Ulangan Harian di Sekolah Dasar dengan Model Rasch. *6*(1), 10-19.

Fikri, S., Machmudah, U., Halimi, H., & Ibrahim, F. M. A. (2021). The Debate Strategy And Its Contribution To The Arabic Learner’s Competence/)3(*4* .استراتيجيّة المناظرة و اسهامها على كفاءة المتعلّم اللغة العربية.

Griffin, P. J., & Nix, P. (1991). *Educational assessment and reporting: A new approach*.

Guntur, W. (2012). Pengaruh person-organization fit, kepuasan kerja dan komitmen organisasi terhadap kinerja perawat. *1*(1), 1-7.

Izza, L. N., Susilaningsih, E., & Harjito, H. (2014). Analisis Instrumen Performance Assessment dengan Metode Generalizability Coefficient pada Penilaian Keterampilan Dasar Laboratorium. *3*(1).

Jimaa, S. (2011). The impact of assessment on students learning. *28*, 718-721.

Linn, R. L. (1991). *Measurement and Evaluation in Teaching*. New York: Macmillan Publishing Company.

Mardapi, D. (2004). Penyusunan tes hasil belajar.

Mardapi, D. (2008). Teknik penyusunan instrumen tes dan nontes. In: Yogyakarta: Mitra Cendikia Press.

Mardapi, D. (2012). Pengukuran penilaian dan evaluasi pendidikan. *45*.

Matt, G. E., Hovell, M. F., Zakarian, J. M., Bernert, J. T., Pirkle, J. L., & Hammond, S. K. (2000). Measuring secondhand smoke exposure in babies: the reliability and validity of mother reports in a sample of low-income families. *19*(3), 232.

Nurbayan, D. R., Nurbayan, Y., & Falah, K. N. (2020). Grammatical Error of Arabic Language in Student Thesis Department of Education Arabic Language FBPS UPI/Kesalahan Nahwu Bahasa Arab Dalam Skripsi Mahasiswa Departemen Pendidikan Bahasa Arab FBPS UPI. *3*(2).

Ramadani, F., & Baroroh, R. U. (2020). Strategies And Methods Of Learning Arabic Vocabulary/Strategi Dan Metode Pembelajaran Kosakata Bahasa Arab. *3*(2).

Retnowati, T. H. (2012). The Development of Assessment Instrument for Elementary School Student Painting. *16*(2), 492-510.

Stiggins, R. J. (1994). *Student-centered classroom assessment*: Merrill New York.

Suskie, L. (2018). *Assessing student learning: A common sense guide*: John Wiley & Sons.

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston: Houghton Mifflin Company.

Williams, D. R., Roggenbuck, J. W., Patterson, M. E., & Watson, A. E. (1992). The variability of user-based social impact standards for wilderness management. *38*(4), 738-756.

Woodward, J. A., & Joe, G. W. (1973). Maximizing the coefficient of generalizability in multi-facet decision studies. *38*(2), 173-181.

Wulansari, A. D., Kumaidi, & Hadi, S. (2019). Two Parameter Logistic Model with Lognormal Response Time for Computer-Based Testing. *14*(15).

Wulansari, A. D., Kumaidi, Hadi, S., Saleh, M., & Friyatmi. (2019). Detection of Students’ Interest With the Logistics Model. *8*(2), 564-571.