Kurdish Studies
Apr 2023
Volume: 11, No: 2, pp. 3906-3920
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

Received: May 2023 Accepted: June 2023
DOI: https://doi.org/10.58262/ks.v11i2.282

Development of Instruments for Measuring Arabic Speaking Ability

Ana Taqwa Wati[1], Trie Hartiti Retnowati[2], Sugeng Sugiyono[3]

Abstract

The study aimed to examine the development of instruments measuring the ability to speak Arabic. This research uses development-based quantitative methods. The development method used is the development method of Rama B. Radhakrishna. The development process is done through five steps: background, conceptualization, format & data analysis, validity, and Reliability. The research subjects included PBA UMY Study Program students who had at least taken lectures. Determination of construct validity using Exploratory Factor Analysis (EFA) technique, for small-scale tests. Confirmatory Factor Analysis (CFA) techniques for large-scale tests. The results of the readability test show that the measurement model of Arabic speaking ability according to experts in general can be understood and can be implemented in lectures of PBA UMY Study Program. The results of validation of model A and model B instruments are obtained that all items for each aspect have a high validity category. The results of limited trials on the measurement instruments of Arabic speaking ability model A and model B obtained information that all task items had a moderate level of difficulty. The results of large-scale trials show that the instrument has acceptable construct validity and Reliability. The conclusion of the results of this study is based on the results of the validity test and Reliability Statistics, it can be concluded that the results of construct validity and Reliability have qualified as good validity and Reliability.

Keywords: *Instrument, Measurement, Speaking Ability, Arabic.*

Introduction

Arabic is an international language because it has been recognized by the United Nations and it is juxtaposed with five other official languages (Spanish, Russian, French, Chinese, and English) (Versteegh, 2014). According to data from The World Factbook (CIA) stated that Arabic is used as the official state language of 27 countries spread across the Middle East. With so many users of this language, it can be assumed that there are also many learners of this language. In addition to the 27 countries where Arabic is an official language, many countries are interested in learning this language (Oueslati et al., 2020). Arabic learners other than Middle Eastern countries, one of which is Indonesia. Indonesians are interested in learning Arabic because the majority of the population is Muslim. Indonesia is also a Muslim-majority country, and with the largest Muslim population worldwide. Based on this, there are many Arabic language learners from Indonesia (Tohe, 2018).

Teaching Arabic in general for Indonesian students is not an easy thing because of the differences between the two languages both in terms of grammar, language culture, pronunciation rules, grammar, and semantics. Students who do not have basic Arabic skills will need more energy to learn Arabic compared to students who already have basic Arabic before(Albantani & Madkur, 2019).

This research will focus on Arabic language competence, especially on speaking skills for PBA UMY

[1]Arabic Language Education, Universitas Negeri Yogyakarta, Indonesia. Email: ana_tw@umy.ac.id
[2]Art Education, Universitas Negeri Yogyakarta, Indonesia. Email: tri_hartiti@yahoo.com
[3]Arabic Literary Language, Yogyakarta State Islamic University, Indonesia. Email: sugeng.sugiyono@uin-suka.ac.id

Study Program students who have studied for at least 3 semesters to find out the results of learning their speaking proficiency in Arabic. The reason for determining the subject is based on the curriculum of PBA UMY Study Program which emphasizes learning Arabic language competence in semesters 1, 2, and 3. Where the distribution of Arabic proficiency material in first semester consists of courses al-istima' wa al-kala>m, naz}ariyyah al-kita>bah and qira>'ah an-nus}u>s}. The distribution of Arabic proficiency material in semester 2 consists of courses al-muna>d}arah, qira>'ah al-kutu>b, tat}bi>q al-kita>bah and 'ilm as}-s}arf. The distribution of Arabic proficiency material in semester 3 consists of 'ilm an-nahw and at-tarjamah courses. The provisional evaluation based on observations and interviews with the head and secretary of the Study Program shows that there is no measuring instrument for Arabic speaking skills for students who have studied for at least 3 semesters in the PBA UMY Study Program.

The number of students is not large, making it impossible to do a placement test for students at the beginning of their studies, but the lecturers hope that there will be an evaluation tool that can measure the level of student ability as a tool to find out the results of student learning outcomes while studying at the UMY PBA Study Program. The problem of this research is that there is no Arabic speaking proficiency measurement tool developed in Indonesia. There is no measuring instrument for Arabic speaking skills for students who have studied for at least 3 semesters in the UMY PBA Study Program. There is no instrument to measure the ability to speak Arabic for students who have studied for at least 3 semesters in the UMY PBA Study Program. It is not yet known how the characteristics of the instrument measuring the ability to speak Arabic for students who have studied for at least 3 semesters in the UMY PBA Study Program. There is no known profile of Arabic speaking skills for students who have studied for at least 3 semesters in the UMY PBA Study Program.

Some similar studies have been conducted by several researchers before. Research conducted by (Zaim et al., 2020) aims to uncover teacher problems in authentic assessment to evaluate English speaking skills of junior high school students, to determine teacher needs in authentic assessment of speaking skills, and to develop authentic assessment models that suit teacher needs and student characteristics. The research conducted by (Ounis, 2017) aims to explore the concept and actual practice of measuring EFL in learners of speaking skills at the tertiary level. Research conducted by (Matrokhim, 2021) The aim is to determine students' self-assessment of Arabic speaking skills. (Zeinoun et al., 2020) aimed to develop an Arabic verbal memory test using a systematic item selection procedure and then provide evidence of validity and Reliability in Arabic-speaking samples in Lebanon.

Based on this previous research, this study offers novelty to focus on developing instruments for measuring Arabic Speaking Ability in this case using the case of PBI UMY Study Program. So, the purpose of this study is to develop instruments that can then be used to measure the ability to speak Arabic accurately and reliably. Finding the instrument construct of the ability to speak Arabic. Acquire quality Arabic speaking instruments. Formulate a profile of Arabic speaking ability.

Theoretical Framework

Arabic Language Skills

Learning Arabic is learning the development of competence in Arabic professionally in all areas of language proficiency. Language proficiency in Arabic is called mahārah lughawiyah which includes four skills, namely: al-Istimā' (listening), al-Kalām (speaking), al-Qira'ah (reading), and al-Kitābah (writing) in meeting personal and social needs (Hanafy, 2014). Speaking is the most important component: people who understand language refers to speakers of that language, as in speaking includes everything related to comprehension and other things, every language learner is more interested in learning speaking skills. four Arabic language skills both in terms of al-Istimā' (listening), al-Kalām (speaking), al-Qira'ah

(reading), and al-Kitābah (writing), thus overall understanding the source of linguistics (your, 2012).

Development Theory

Research and Development (R &; D) or research and development is a process or steps to develop a new product or perfect an existing product, which can be accounted for (Barge-Gil &; López, 2014; Mehand et al., 2018). Research and development is a method commonly used in research to obtain a product, and test the effectiveness of the product. Based on the two opinions above, it can be concluded that a study is research that can produce new products, based on the products to be developed (Sugiyono, 2013).

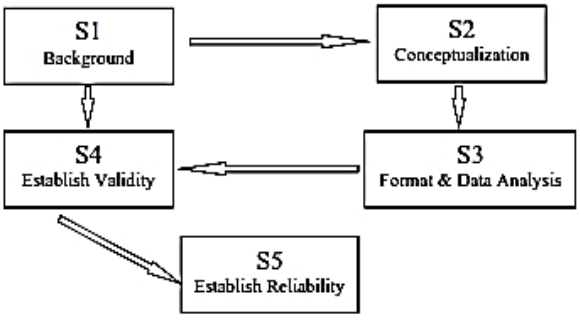
In vocational and adult education, there is a new need to create a workforce with flexibility and concern for accountability of educational outcomes, so that many people are very hopeful that the end results of learning will prove to be concrete, practical and relevant skills. To obtain output knowledge in accordance with these expectations, the study program must have a measuring tool for the ability of its students at the end of their study period (Fulcher, 2014). This measuring tool is in the form of an evaluation tool for Arabic language skills for all students of PBA UMY Study Program.

Research Methods

Development model

The development model to be used in this research is the Rama B. Radhakrishna development model. This model was outlined in a paper published in the Journal of Extension, Knowledge Sharing in 2007. The choice of this development model is because in its development technique, this method can answer and meet the research needs of developing instruments for measuring Arabic speaking skills for UMY PBA Study Program students. Questions and needs in this study include how to produce instrument constructs, know the characteristics of floating instruments, and know the validity and Reliability of the instruments developed. The development scheme to be carried out is as illustrated in the chart as follows:

Figure 1: Instrument Development Flowchart.



The first step (S1) is the background problem of why researchers feel the need to develop this instrument; The second step (S2) is the concept of the instrument to be developed; The third step (S3) is the development format and data analyzed; The fourth step (S4) establishes the validity of the developed instrument, and the fifth step (S5) establishes the Reliability of the developed instrument.

Try Subjects

The test subjects in this development research are PBA UMY study program students who are considered to have mastered Arabic speaking competence, namely all students who have taken Arabic speaking competency courses. Students who are referred to in this study are students who have at least taken all Arabic language competency courses or at least 3 semesters. The subjects were selected from

all PBA UMY Study Program students who had mastered the competence to speak Arabic because if the researcher only took seventh semester students, the number of students was too small. The object was chosen referring to the curriculum of PBA UMY Study Program. PBA UMY Study Program places all courses containing Arabic language competence in semesters 1, 2, and 3. Thus, it can be assumed that students who have studied for at least 3 semesters have Arabic language competence.

Data Collection Techniques and Instruments

Teknik dokumentasi digunakan dalam penelitian ini dikarenakan dibutuhkannya data-data berdasar pada dokumen yang dimiliki Prodi PBA UMY, serta beberapa dokumen dari Fakultas ISIPOL dan fakultas Pendidikan Bahasa UMY. Focus Group Discussion (FGD) adalah teknik pengumpulan data yang umumnya dilakukan pada penelitian kualitatif dengan tujuan menemukan makna sebuah tema menurut pemahaman sebuah kelompok. Teknik ini digunakan untuk mengungkap pemaknaan dari suatu kalompok berdasarkan hasil diskusi yang terpusat pada suatu permasalahan tertentu. FGD juga dimaksudkan untuk menghindari pemaknaan yang salah dari seorang peneliti terhadap fokus masalah yang sedang diteliti. Forum ini diikut oleh dosen bahasa Arab di Prodi PBA UMY sebagai pelaksana lapangan, Ahli Bahasa Arab, dan Ahli Pengembangan Instrumen. Interview. Menurut Sugiyono (2006) wawancara dapat dilakukan secara terstruktur dan tidak terstruktut, dan dapat dilakukan dengan tatap muka (face to face) maupun menggunakan telepon. Interview digunakan dalam penelitian pengembangan ini karena peneliti membutuhkan beberapa data yang hanya dapat diperoleh melalui interview. Interview dilakukan oleh peneliti kepada Dekan Fakultas ISIPOL dan Fakultas Pendidikan Bahasa, serta Ketua dan Sekertaris Prodi PBA UMY, tenaga kependidikan, juga mahasiswa Prodi PBA UMY. Angket dalam penelitian ini berupa pertanyaan-pertanyaan yang diberikan kepada ahli, dan mahasiswa yang berfungsi sebagai rater bagi instrumen yang dikembangkan. Masukan, kritik, serta saran dijadikan sebagai acuan merevisi produk yang dikembangkan. Hasil revisi berupa produk instrumen yang siap untuk diuji cobakan.

Data Analysis Techniques

The data analysis techniques used are adjusted to the type of data collected. Some things to consider in data analysis include: 1.) data analysis includes procedures for organization, reduction, and presentation of data, whether with tables, charts, or graphs; 2.) data are classified by type and components of developed products; 3.) the data is analyzed descriptively as well as in the form of quantitative calculations; 4.) the presentation of data analysis results is limited to factual matters, with no developer interpretation, so as a basis for conducting and revising the model; and 5.) Data analysis uses calculations and statistical analysis in line with the problems posed and products developed.

Based on the type of instrument used, this study produces two types of data, namely qualitative data and quantitative data. The resulting qualitative data is basically in the form of preliminary study data and model feasibility data. The quantitative data produced in this study are data related to the hypothesis proposed. Quantitative data is analyzed using Aiken's V formula to view instrument items based on expert judgment. Data analysis in small-scale trials using confirmatory factor analysis to see empirical content validity. Large-scale trial stage data analysis also uses confirmatory factor analysis to see construct validity. The effectiveness of the model was analyzed using descriptive statistics by creating categories of 3.26 to 4.00 with very good or very effective categories; 2.51 to 3.25 with good or effective category: 1.76 to 2.50 with good or effective category; and 1 to 1.75 with poor or ineffective categories (Sultan Nurhadı, Endah TriPrıyatnı4, 2017). Effectiveness is assessed in part using experts or practitioners who have used the instrument, as no previous instrument has been developed to evaluate Arabic speaking ability. This method can be used if no model, instrument or product has been previously developed to assess a program (Morrison-Saunders et al., 2021; Pope et al., 2013).

Result

Initial Draft Development

The development of the initial draft of the model is carried out through the development of the initial draft of the model (*prototype*) and model validation. The development of the draft model aims to produce a model that can measure the ability to speak Arabic in students well, can be implemented, and is effectively used in measuring learning outcomes in lectures. The draft model for measuring Arabic speaking skills for students is carried out through model *prototype* design activities which include model objectives, model characteristics, model components, model instruments, model syntax, and model implementation guidelines. Furthermore, this initial draft is reviewed by supervisors and experts in the field of language and measurement/evaluation experts.

Referring to the supervisor's review, several suggestions / inputs were obtained regarding the initial draft of the model. Suggestions / inputs from supervisors and experts include: (1) the objectives of the model must be clearly prepared and in accordance with the principles of skill assessment; (2) the characteristics of the model are made specific and clear; (3) model components and model instruments need to be clarified, scrutinized, and improved so that they are clearer and more focused; (4) the syntax of the assessment model is prepared more operationally and systematically; and (5) the model implementation guide is prepared more simply, practically, and completely.

The components of the model assessed by experts include material aspects, construction aspects, and language aspects. The model validation was carried out by five experts, namely 2 (two) linguists/PBAs and 3 (three) measurement/evaluation experts. The results of validation against the model are presented in Table 1. and Table 2.

Table 1: Results of Analysis Validation of the Initial Draft of the Model.

No	Instruments	Expert Advice
1.	Arabic Speaking Ability Measurement Model	In general, it is worth using Need revision: (1) consistency of definitions and concepts of speaking ability and (2) syntax clarified, Model usage guidelines need to be simpler, more operational, and more complete.
2.	Arabic Speaking Instrument Model **A**	1. Commands need to be clarified 2. Use of context-tailored diction 3. The question items are worth using after going through revision
3.	Arabic Speaking Instrument Model **B**	

Table 2: Summary of Prototype Model of Instrument Measuring Arabic Speaking Ability.

No.	***Prototype* Model**	Description
1	Model Purpose	To measure/assess the speaking ability of Arabic learner students.
2	Model Features	1. Integrated with lectures 2. The assessment focuses on the ability to speak Arabic 3. The instrument used is in the form of performance appraisal. 4. Lecturers can give assignments in the form of exercises/practices to develop students' ability to speak Arabic.
3	Model Components	1. Measured competency 2. Model A and model B speaking ability assessment instruments, which contain tasks, scoring guidelines and assessment rubrics. 3. Model guide 4. Data on assessment results 5. Report on assessment results
4	Model Instruments	Speaking ability tasks, scoring guidelines, and assessment rubrics (Model A and Model B)
5	Model Syntax	1. Instilling concepts, principles and applications of various methods oriented towards life skills 2. Application of the concepts of skills and beauty of spoken language in international relations Study of the implications of developing or implementing science and technology in problem solving
6	Model Guide	1. Model implementation 2. Scoring and grading techniques 3. Reporting research results 4. Utilization of Assessment results

Results of the Development of Instruments for Measuring Arabic Speaking AbilityThe procedure for developing an instrument model for measuring Arabic speaking ability uses a type of performance assessment which includes two instrument models, namely model A and model B. Model A is an instrument for measuring Arabic speaking ability using a face to face interview model between students and examining lecturers, while model B is an instrument measuring Arabic speaking ability which requires students to make presentations in front of classmates who are assessed by the lecturer. The development of Arabic speaking ability instruments refers to the competency standards involved in its design.

The results of developing an instrument for measuring Arabic speaking ability through design, as a result of reviewing and improving the test/performance items, are presented in Table 3.

Tabel 3: Hasil Revisi Instrumen Penilaian

No	Instrumen	Number of Items Previously	Fixed	Number of Items After Validation
1	Model A	15	3	15
2	Model B	7	2	7

Hasil Uji Coba Produk

The product trial began with a readability test on the model A and model B Arabic speaking ability measurement instruments carried out by lecturers and students. The readability test was carried out by giving a readability questionnaire to Arabic language lecturers and students, followed by interviews to gather information regarding responses to the questionnaire. The readability test activity involved a lecturer and 3 (three) UMY PBA Study Program students. The readability test results are used to reflect and revise for the next trial. Table 4 presents the results of the readability test on the model and assessment instruments.

Table 4: Readability Test Results

No	Instrumen	Saran
1.	Instrument Model for Measuring Arabic Speaking Ability	1. In general, the model can be understood and implemented in PBA lectures. 2. The terminology is simplified to make it easier to understand. 3. Procedures are more simplified and operational
2.	Model A Arabic Speaking Ability Measurement Instrument	1. Assignments are in accordance with lecture material. 2. The assignment model is different from what is usually given.
3.	Model B Arabic Speaking Ability Measurement Instrument	3. It is necessary to prepare an assessment rubric to measure speaking proficiency achievements

The results of validation by experts on the Arabic speaking ability measurement instruments model A and model B are shown in tables 5 and 6.

Table 5: Validation results for model A instrument

No.	Aspek					
item	Material	Category	Construction	Kategori	Construction	Category
1	1,00	High	0,89	High	1,00	High
2	1,00	High	0,89	High	0,89	High
3	1,00	High	1,00	High	1,00	High
4	1,00	High	1,00	High	1,00	High
5	1,00	High	0,89	High	0,89	High
6	1,00	High	0,89	High	1,00	High
7	1,00	High	1,00	High	1,00	High
8	1,00	High	0,89	High	0,89	High
9	1,00	High	1,00	High	1,00	High
10	1,00	High	0,89	High	0,89	High
11	1,00	High	1,00	High	1,00	High
12	1,00	High	0,89	High	0,89	High
13	1,00	High	0,89	High	0,89	High

14	1,00	High	1,00	High	1,00	High
15	1,00	High	0,89	High	1,00	High
Prkt	1,00	High	0,93	High	0,96	High

Based on the results in Table 7, information is obtained that all items for each aspect have a high validity category, because the lowest Aiken index is 0.89, the highest is 1.00 or more than 0.8. According to Heri Retnawati (2016: 38), items that have an index of less than 0.4 are said to have low validity, an index in the range of 0.4 - 0.8 is said to have moderate validity, and if the index is more than 0.8 then the validity is high. Likewise, the index for the instrument, in all aspects, also has a high validity category. Thus, it can be stated that the model A Arabic speaking ability measurement instrument can be applied in the next trial

Table 6: Validation Results for Model B Instruments

No. butir	Aspect					
	Material	**Category**	**Construction**	**Category**	**Language**	**Category**
1	0,89	High	0,89	High	1,00	High
2	1,00	High	0,89	High	0,89	High
3	0,89	High	0,89	High	0,89	High
4	1,00	High	0,89	High	0,89	High
5	1,00	High	1,00	High	1,00	High
6	1,00	High	0,89	High	0,89	High
7	1,00	High	0,89	High	0,89	High
Prkt	0,97	High	0,90	High	0,92	High

The data in Table 6 explains that all items for each aspect (material, construction and language) have a high validity category, because the lowest Aiken index is 0.89, the highest is 1.00 or more than 0.8. Likewise, the index for the instrument, in all aspects, also has a high validity category. Thus, it can be stated that the Arabic speaking ability measurement instrument model B can be applied in the next trial.

Limited Trial of the Instrument for Measuring Arabic Speaking Ability

A limited trial was carried out with the aim of determining the suitability of the item model with the Rasch model, the criteria for whether or not the items were accepted, and the difficulty index of the test items/tasks on the instrument measuring Arabic speaking ability. The trial subjects were limited to 32 students for the model A instrument and 65 students for the model B instrument. The trial result data was then analyzed using the QUEST program.

Limited Trial of Model A Arabic Speaking Ability Measurement Instrument

The results of the model suitability test using the QUEST program showed that there were 32 subjects analyzed with 15 test items/tasks at a probability of 0.5 in accordance with the Maximum Likelihood principle. From the test results, information was obtained that the mean INFIT MNSQ was 0.96 and the SD was 0.86, meaning that overall the test items/tasks were in accordance with the Rasch model. In addition, the higher the value, the more convincing it is that the measurement provides consistent results.

The criteria for whether a question item is accepted or not can be determined by looking at the results of the INFIT MNSQ analysis. INFIT t calculation with a limit of ±2.0 shows that all assignment items are accepted. At the end of the analysis, an internal consistency value of 0.62 is presented, which means it shows a high reliability value.

The task item difficulty index in analysis with the QUEST program is obtained by looking at the delta value.

The results of calculating item quality using the item response theory approach are presented in Table 7.

Table 7: Results of Model A Instrument Item Quality Analysis

No	Fit to the Rasch Model			Difficulty Index (delta value)	
Item	Infit Meansquare	Infit t	Information	Value	Information
1	0,78	-0,8	Accepted	0,31	Medium
2	0,82	-0,6	Accepted	0,37	Medium
3	0,67	-1,1	Accepted	0,53	Medium
4	0,85	-0,3	Accepted	0,52	Medium
5	0,98	0,0	Accepted	0,50	Medium
6	0,90	-0,3	Accepted	0,39	Medium
7	0,81	-0,6	Accepted	0,37	Medium
8	1,66	2,0	Accepted	0,34	Medium
9	1,08	0,3	Accepted	0,42	Medium
10	1,17	0,6	Accepted	0,36	Medium
11	0,94	-0,1	Accepted	0,40	Medium
12	1,02	0,2	Accepted	0,56	Medium
13	1,04	0,2	Accepted	0,44	Medium
14	0,70	-0,9	Accepted	0,36	Medium
15	0,96	-0,1	Accepted	0,31	Medium

Refer to Table 7. It was obtained that from 15 items of questions / tasks to measure Arabic speaking proficiency, statistically showed that the INFIT t score with a limit of ±2.0. Thus all assignment items are accepted. Furthermore, information is obtained that all task items have a 'medium' difficulty level.

Overall each task item meets the criteria as good, therefore these fifteen task items are acceptable and worthy of use in expanded trials. Meanwhile, the calculation of the reliability coefficient of the Arabic speaking proficiency task showed a number of 0.62. The amount of coefficiency can already be said to be good for model development research.

Limited Trial of Model B Arabic Speaking Measurement Instrument

The results of the model fit test using the *QUEST* program showed that there were 65 subjects analyzed with 7 test items/tasks at odds of 0.5 in accordance with the principle of *Likelihood Maximum*. From the test results, information was obtained that the mean *INFIT MNSQ* 1.00 and SD 1.50 means that the overall test items / tasks are in accordance with the *Rasch* model. In addition, the higher the value the more reassuring that the measurement gives consistent results.

The criteria for whether or not the question items are accepted can be determined by looking at the results of the INFIT MNSQ analysis. The calculation of INFIT t with a limit of ±2.0 shows that all assignment items are accepted. At the end of the analysis, an internal consistency value of 0.52 is presented, which means that it shows a fairly high reliability value.

The difficulty index of task items in analysis with the QUEST program is obtained by looking at the delta value. The results of the calculation of grain quality calculations with the grain response theory approach are presented in Table 8.

Table 8: Results of Model B Instrument Item Quality Analysis

No	Fit to the Rasch Model			Difficulty Index (delta value)	
Item	Infit Meansquare	Infit t	Ket.	Value	Ket
1	1,08	0,5	Accepted	0,29	Medium
2	1,13	0,5	Accepted	0,32	Medium
3	1,13	0,8	Accepted	0,25	Medium
4	0,96	-0,1	Accepted	0,25	Medium

5	0,95	0,0	Accepted	0,54	Medium
6	0,84	-1,0	Accepted	0,21	Medium
7	0,90	-0,5	Accepted	0,34	Medium

Referring to Table 8, information is obtained that from the 7 questions/tasks for the instrument measuring Arabic speaking ability, statistically it shows that the INFIT t value is within ±2.0. Thus all assignment items are accepted. Furthermore, information was obtained that all task items had a 'medium' level of difficulty. Overall, each task item meets the criteria for being a good item. Therefore, these eight task items are acceptable and suitable for use in expanded trials. Meanwhile, the calculation of the task reliability coefficient on the instrument measuring Arabic speaking ability shows the number 0.52. The magnitude of this coefficient can be said to be good enough for model development research.

Large scale trials

Model A Instrument

Large-scale trials are carried out to see the validity and Reliability of the constructs obtained from the factors or components to be evaluated. The constructs or indicators obtained from the factors will be seen for their validity and Reliability. This was done to prove that the constructs found through FGD and literature review were valid and reliable. Validity and Reliability for the model A instrument was carried out on 120 respondents and analyzed with CFA using Lisrel 8.80 software. The detailed analysis results can be seen in Figure 2 and Table 9.

Figure 2: CFA Results Using The LISREL 8.80 Program.

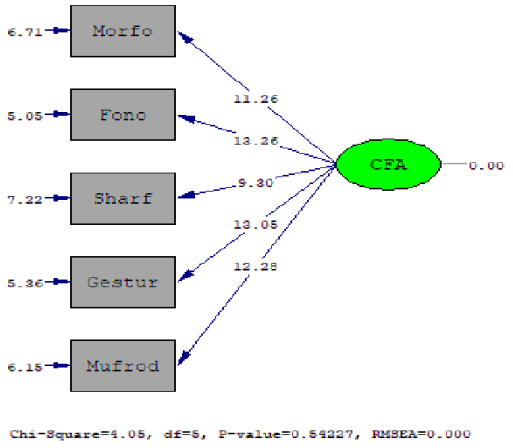


Table 9: Model Fit Criteria.

Goodness of fit index	Criteria	Result	Status
Empirical chi-square	Chi square < 2 db	4.05 <2 (5)	OK
Root Mean Square Approximation (RMSEA)	≤ 0.08	0,000	OK
P-Value	>0.05	0.5422	OK
GFI	>0.90	0.96	OK
AGFI	>0.90	0,99	OK
NFI	>0.90	0,99	OK

The results of the analysis in Table 9 explain that the construct factors morphology/s{arf, phonology/makha>rij al-h{arf, syntax/ qaidah, gestures & expressions, and vocabulary/mufrada>t have acceptable validity values and can be used to evaluate Arabic speaking skills for students studying Arabic at universities. From the analysis results obtained Chi-Square = 4.05 < 2(5), P-Value = 0.5422, Root Mean Square Effort of Measurement (RMSEA) = 0.000, Goodness of Fit Index = 0.96 and Adjusted Goodness of Fit Index (AGFI) = 0.99, Normed Fit Index (NFI) = 0.99. Based on the

Goodness of Fit value, it can be concluded that the data obtained is in accordance with the measurements developed. Construct validity and Reliability can be seen in Table 10 and Table 11.

Table 10: Summary of Construct Validity.

No	Indicator	Loading > 0,4 (standardize)	Loading>1.96 T-Value	Criteria
1	Morphology / *s{arf*	0,84	11.26	Valid
2	Phonology / *makha>rij al- h{{arf*	0,93	13.26	Valid
3	Syntax / *qaidah*	0,74	9.30	Valid
4	Gestures and expressions	0,92	13.05	Valid
5	Vocabulary / *mufrada*>t.	0.89	13.28	Valid

Table 11: Summary of Construct Reliability.

No	Indicator	Loading	Error	CR	Conclusion
1	Morfologi/s{arf	0.84	0.29	0.94	Reliabel
2	Phonology/makha>rij al- h{{arf	0.93	0.14		
3	Syntax/qaidah	0.74	0.45		
4	Gestures and expressions	0.92	0.15		
5	Vocabulary/mufrada>t.	0.89	0.21		

Based on Table 10 and Table 11, validity and Reliability statistics can be concluded that the results of construct validity and Reliability meet the requirements as good validity and Reliability. Construct validity has exceeded the standard for the factors morphology/s{arf, phonology/makha>rij al-h{arf, syntax/qaidah, gestures & expressions, and vocabulary/mufrada>t, namely 0.5, while construct reliability is greater than 0.7. The validity and reliability results obtained from Tables 12 and 13 are the results of running the LISREL 8.80 program and show that instrument, process and output resources can be used to evaluate Arabic speaking skills for Arabic language students.

Model B Instrument

This reliability test was calculated to determine the amount of Reliability between raters in assessing Arabic speaking ability. This approach is used to assess agreement between raters in assessing an individual. Thus, this Reliability is attached to the score given, not to the measuring instrument. This trial was carried out on 120 students by means of students giving presentations that they had prepared beforehand, and assessed by 2 lecturers. The results of the ICC (Intraclass Correlation Coefficients) calculation for the Arabic speaking ability measurement instrument were obtained:

Table 12: Alpha Cronbach of Arabic speaking ability model B.

Reliability Statistics	
Cronbach's Alpha	N of Items
.973	2

Based on table 12 Reliability Statistics for estimation using Cronbach's Alpha coefficient, it shows that the Alpha reliability of the instrument measuring Arabic speaking ability model B is 0.973. This figure is a very large number so it can be said that the Alpha reliability for the instrument measuring Arabic speaking ability model B is very reliable.

Table 13: ANOVA of Arabic speaking ability model B.

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig
Between People	145.793	119	1.225		

Within People	Between Items	.000	1	.000	.008	.930
	Residual	3.953	119	.033		
	Total	3.953	120	.033		
Total		149.746	239	.627		
Grand Mean = 3.5583						

Table 13 is to see whether or not there are differences in assessments between raters. Based on the table above, it can be seen that p>0.05 or 0.930>0.05, which indicates that there is no difference in assessments between raters. The ICC reliability coefficient values are presented in table 14 as follows:

Table 14: ICC reliability coefficient values between raters for Arabic speaking ability model B

Intraclass Correlation Coefficient							
	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.947[a]	.925	.963	36.883	119	119	.000
Average Measures	.973[c]	.961	.981	36.883	119	119	.000

Based on the table above, the intraclass correlation coefficient or inter-rater reliability value is rxx = 0.947. The value of 0.947 is included in the large reliability coefficient category.

Discussion

The development of the initial draft of the model is carried out through the activities of developing an initial draft of the model (prototype) and model validation. Next, this initial draft is reviewed by supervisors and experts in the field of language and measurement/evaluation experts. The results of the validation analysis of the initial draft of the model are in the form of expert advice. Expert advice obtained includes; First, the instrument model for measuring Arabic speaking ability. In general, it is suitable for use. Revision is needed: (a) Consistency of definitions and concepts of speaking proficiency and (b) syntax is clarified. Guidelines for using the model need to be prepared to be simpler, more operational and complete. Second, the instrument for measuring Arabic speaking ability Model A & B, commands need to be clarified, use of diction that is adapted to the context., question items are suitable for use after going through revision.

The summary of the prototype instrument model for measuring Arabic speaking ability that can be presented in the discussion is:

a. The aim of the model is to measure or assess the speaking abilities of Arabic language students.
b. The characteristics of the model are integrated with lectures, the assessment focuses on the ability to speak Arabic, the instruments used are performance assessments, and lecturers can give assignments in the form of exercises/practices to develop students' ability to speak Arabic.
c. The model components are measured model components, model A and model B Arabic speaking ability measurement instruments which contain tasks, scoring guidelines and assessment rubrics, model guidelines, assessment result data and assessment result reports.
d. The model instruments are in the form of Arabic speaking ability measurement instruments, scoring guidelines, and assessment rubrics (Model A and Model B)
e. The model syntax is in the form of embedding concepts, principles and application of various methods oriented towards life skills, application of skills concepts and the beauty of spoken language in international relations, studying the implications of developing or implementing science and technology in solving problems.
f. The model guide consists of implementing the model, scoring and assessment techniques, reporting research results, and utilizing assessment results.

The draft results of the development of the Arabic speaking instrument model include two instrument models, namely the Arabic speaking ability measurement instrument model A and the Arabic speaking ability measurement instrument model B. Model A instrument is an instrument measuring the ability to speak Arabic with a face to face interview model between students and examining lecturers, while model B instruments are instruments that demand student presentations in front of classmates assessed by lecturers. The results of the revision of the assessment instrument for each model can be reported that the model A instrument consists of 15 question items, there are 3 questions that need to be corrected, 0 questions are discarded, and the number of items after validation is 15 questions. The assessment of the revised results of the model B instrument assessment instrument consists of 7 question items, there are 2 questions that need to be corrected, 0 questions are discarded, and the number of items after validation is 7 questions.

The draft results of the development of the Arabic speaking instrument model include two instrument models, namely the Arabic speaking ability measurement instrument model A and the Arabic speaking ability measurement instrument model B. Model A instrument is an instrument measuring the ability to speak Arabic with a face to face interview model between students and examining lecturers, while model B instruments are instruments that demand student presentations in front of classmates assessed by lecturers. The results of the revision of the assessment instrument for each model can be reported that the model A instrument consists of 15 question items, there are 3 questions that need to be corrected, 0 questions are discarded, and the number of items after validation is 15 questions. The assessment of the revised results of the model B instrument assessment instrument consists of 7 question items, there are 2 questions that need to be corrected, 0 questions are discarded, and the number of items after validation is 7 questions.

Test results on measuring the ability to speak Arabic model A and measuring instruments for the ability to speak Arabic model B conducted by lecturers and students. In the product trial stage, researchers provided readability questionnaires to Arabic lecturers and students, followed by interviews to explore information about responses to questionnaires. The results of the instrument readability test get several suggestions, namely: in general, the model can be understood and implemented in lectures, terminology is simplified to make it easy to understand, and procedures are more simplified and operational. The results of the readability test of the measurement instrument of the ability to speak Arabic model **A** and the measurement instrument of the ability to speak Arabic model **B** received several suggestions, namely: the assignment is in accordance with the lecture material, the assignment model is different from what is usually given, and it is necessary to prepare an assessment rubric to measure the achievement of speaking ability.

The results of instrument validation by experts on both instruments measuring the ability to speak Arabic obtained information that all items for each aspect have a high validity category, because the lowest Aiken index of 0.89 is the highest of 1.00 or more than 0.8. According to Heri Retnawati (2016: 38) items that have an index of less than 0.4 are said to have low validity, an index in the range of 0.4 – 0.8 is said to be of medium validity, and if the index is more than 0.8 then the validity is high.

Limited trial of the Model A Arabic speaking instrument. The results of the model fit test using the QUEST program showed that there were 32 subjects analyzed with 15 test items/tasks at odds of 0.5 in accordance with the principle of Likelihood Maximum. From the test results, information was obtained that the mean INFIT MNSQ 0.96 and SD 0.86 means that overall test items / tasks are in accordance with the Rasch model. Based on the results of the previous analysis, information was obtained that from 15 items / tasks to measure the ability to speak Arabic, statistically showed that the INFIT value t with a limit of ±2.0. Thus all assignment items are accepted. Furthermore, information is obtained that all task items have a 'medium' difficulty level.

Limited trial of the Model B Arabic Speaking Measurement Instrument. The results of the model fit test using the QUEST program showed that there were 65 subjects analyzed with 7 test items/tasks at odds

of 0.5 in accordance with the principle of Likelihood Maximum. From the test results, information was obtained that the mean INFIT MNSQ 1.00 and SD 1.50 means that the overall test items / tasks are in accordance with the Rasch model. Referring to the results of the analysis, information was obtained that from 7 items / tasks to measure the ability to speak Arabic, statistically showed that the INFIT value t with a limit of ±2.0. Thus all assignment items are accepted. Furthermore, information is obtained that all task items have a 'medium' difficulty level.

Hasil analisis uji coba produk skala besar diperoleh informasi bahwa konstruk faktor morfologi/qaidah, fonologi/makha>rij al- h{arf, sintaksis/s{arf, gestur & ekspresi, dan kosa kata/mufrada>t memiliki nilai validitas yang dapat diterima dan dapat digunakan untuk mengevaluasi kemampuan berbicara bahasa Arab bagi mahasiswa pebelajar bahasa Arab di universitas. Dari hasil analisis diperoleh nilai Chi-Square = 4,05 < 2(5), P-Value = 0,5422, Root Mean Square Effort of Measurement (RMSEA) = 0,000, Goodness of Fit Index = 0,96 dan Adjusted Goodness of Fit Index (AGFI)) = 0,99, Normed Fit Index (NFI) = 0,99. Berdasarkan nilai Goodness of Fit dapat disimpulkan bahwa data yang diperoleh sesuai dengan pengukuran yang dikembangkan.

Based on the results of the validity test and Reliability Statistics, it can be concluded that the results of construct validity and Reliability have qualified as good validity and Reliability. The construct validity has exceeded the standard for morphological factors / s {arf, phonology / makha >rij al-h {arf, syntax / qaidah, gestures & expressions, and vocabulary / mufrada >t which is 0.5 while the Reliability of the construct is greater 0.7. The results of validity and Reliability obtained are the results of running the LISREL 8.80 program and show that instrument, process, and output resources can be used to evaluate Arabic speaking skills for Arabic language learners.

An instrument developed to evaluate Arabic speaking skills for university students who have studied Arabic at least or taken Arabic courses. The instrument is complete, easy to use, accurate in obtaining information about the weaknesses and advantages of evaluating Arabic speaking skills for students in universities who have at least studied Arabic or taken Arabic courses, and is very useful for universities that have Arabic literature study programs or PBA study programs. Effective instruments will make it easier for users to access information that hinders the success of a program. Effective instruments can describe the components that need to be improved so as to make a strong contribution to improving the developed program. Valid, reliable, and proven effective instruments will show measurement results, so that stakeholders can correct educational program deficiencies.

Conclusion

The results of this development research are in the form of measuring Arabic speaking skills for Arabic learning students. The domains measured in this Arabic speaking ability measurement instrument are morphology/qaidah, phonology/makha>rij al-h{arf, syntax/s{arf, gestures & expressions, and vocabulary/mufrada>t. Based on data analysis, information was obtained that all items for each aspect have a high validity category, because the lowest Aiken index of 0.89, the highest of 1.00 or more than 0.8. In Heri Retnawati's opinion that items that have an index of less than 0.4 are said to have low validity, an index in the range of 0.4 – 0.8 is said to be of medium validity, and if the index is more than 0.8 then the validity is high.

The results of the model suitability test for the model A Arabic speaking ability measurement instrument using the QUEST program showed that there were 32 subjects analyzed with 15 test items/tasks at odds of 0.5 in accordance with the principle of Likelihood Maximum. From the test results, information was obtained that the mean INFIT MNSQ 0.96 and SD 0.86 means that overall test items / tasks are in accordance with the Rasch model. Based on the results of the previous analysis, information was obtained that from 15 items/tasks to measure the ability to speak Arabic, statistically showed that the

INFIT value t with a limit of ±2.0. Thus all assignment items are accepted. Furthermore, information is obtained that all task items have a 'medium' difficulty level.

The results of the model suitability test for the model B Arabic speaking measurement instrument using the QUEST program showed that there were 65 subjects analyzed with 7 test items/tasks at odds of 0.5 in accordance with the principle of Likelihood Maximum. From the test results, information was obtained that the mean INFIT MNSQ 1.00 and SD 1.50 means that the overall test items/tasks are in accordance with the Rasch model. Referring to the results of the analysis, information was obtained that from 7 items/tasks to measure the ability to speak Arabic, statistically showed that the INFIT value t with a limit of ±2.0. Thus all assignment items are accepted. Furthermore, information is obtained that all task items have a 'medium' difficulty level.

The results of large-scale product trial analysis obtained information that the construct of morphological factors/s{arf, phonology/ makha >rij al-h {arf, syntax/qaidah, gestures & expressions, and vocabulary / mufrada>t have acceptable validity values and can be used to evaluate Arabic speaking skills for Arabic learning students at universities. From the results of the analysis obtained Chi-Square value = 4.05 < 2(5), P-Value = 0.5422, Root Mean Square Effort of Measurement (RMSEA) = 0.000, Goodness of Fit Index = 0.96 and Adjusted Goodness of Fit Index (AGFI) = 0.99, Normed Fit Index (NFI) = 0.99. Based on the Goodness of Fit value, it can be concluded that the data obtained correspond to the developed measurements.

Based on the results of the validity test and Reliability Statistics, it can be concluded that the results of construct validity and Reliability have qualified as good validity and Reliability. The construct validity has exceeded the standard for morphology/qaidah, phonology/makharij al-harf, syntax/sharf, gestures & expressions, and vocabulary/Mufradhat which is 0.5 while the construct reliability is greater 0.7. The results of validity and Reliability obtained are the results of running the LISREL 8.80 program and show that instrument, process, and output resources can be used to evaluate Arabic speaking skills for Arabic language learners.

Based on the results of this development research, theoretical and practical implications can be put forward based on the findings in the process of conducting research and the results as follows: Theoretical Implications, The development of language proficiency instruments, especially Arabic, is still needed. The results of this development will be needed and very helpful for lecturers to know the Arabic language skills of their students. The development of good instruments can produce a tool for measuring Arabic language skills well as well. Instruments that are validly and reliably tested can be used as a trusted measuring instrument. Student confidence increases when their abilities have been tested with instruments with a high level of Reliability. The results of this development can test students' abilities, and the results of the assessment can be accounted for. Then, Practical Implications, the results of this development research are the initial draft of instruments measuring Arabic speaking ability. This draft that has been tested and validated by experts and has been tested for its efficacy can be used for Arabic teaching lecturers to measure the ability to speak Arabic for their students.

References

Albantani, A. M., & Madkur, A. (2019). Teaching Arabic in the era of Industrial Revolution 4.0 in Indonesia: Challenges and opportunities. *ASEAN Journal of Community Engagement*, *3*(2), 3. https://doi.org/10.7454/ajce.v3i2.1063

Barge-Gil, A., & López, A. (2014). R&D determinants: Accounting for the differences between research and development. *Research Policy*, *43*(9), 1634–1648. https://doi.org/https://doi.org/10.1016/j.respol.2014.04.017

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

Hanafy, M. S. (2014). Konsep Belajar Dan Pembelajaran. *Lentera Pendidikan : Jurnal Ilmu Tarbiyah Dan Keguruan*, *17*(1), 66–79. https://doi.org/10.24252/lp.2014v17n1a5

Matrokhim, M. (2021). Students' Self-Assessment of Arabic Speaking Skill. *International Journal of Arabic Language Teaching*, *3*(02), 185–195. https://doi.org/https://doi.org/10.32332/ijalt.v3i02.4208

Mehand, M. S., Al-Shorbaji, F., Millett, P., & Murgue, B. (2018). The WHO R&D Blueprint: 2018 review of emerging infectious diseases requiring urgent research and development efforts. *Antiviral Research*, *159*, 63–67. https://doi.org/https://doi.org/10.1016/j.antiviral.2018.09.009

Morrison-Saunders, A., Arts, J., Bond, A., Pope, J., & Retief, F. (2021). Reflecting on, and revising, international best practice principles for EIA follow-up. *Environmental Impact Assessment Review*, *89*, 106596. https://doi.org/https://doi.org/10.1016/j.eiar.2021.106596

Oueslati, O., Cambria, E., HajHmida, M. Ben, & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*, *112*, 408–430. https://doi.org/https://doi.org/10.1016/j.future.2020.05.034

Ounis, A. (2017). The Assessment of Speaking Skills at the Tertiary Level. *International Journal of English Linguistics*, *7*, 95. https://api.semanticscholar.org/CorpusID:157456839

Pope, J., Bond, A., Morrison-Saunders, A., & Retief, F. (2013). Advancing the theory and practice of impact assessment: Setting the research agenda. *Environmental Impact Assessment Review*, *41*, 1–9. https://doi.org/https://doi.org/10.1016/j.eiar.2013.01.008

Sugiyono, D. (2013). *Metode penelitian pendidikan pendekatan kuantitatif, kualitatif dan R&D*. Alfabeta.

Sultan Nurhadı, Endah TriPrıyatnı4, A. (2017). The Effect of the Critical Literacy Approach on Pre-service Language Teachers' Critical Reading Skills. *Journal*, *17*(71), 159–174.

Tohe, A. (2018). Arabic Language At the Crossroad : a Case Study in Indonesia. *Prosiding Pertemuan Ilmiah Internasional Bahasa Arab*, *0*(0), 977–988. http://prosiding.imla.or.id/index.php/pinba/article/view/83

Ur, P. (2012). *A course in English language teaching*. Cambridge University Press.

Versteegh, K. (2014). *Arabic language*. Edinburgh University Press.

Zaim, M., Refnaldi, & Arsyad, S. (2020). Authentic assessment for speaking skills: Problem and solution for english secondary school teachers in Indonesia. *International Journal of Instruction*, *13*(3), 587–604. https://doi.org/10.29333/iji.2020.13340a

Zeinoun, P., Farran, N., Khoury, S. J., & Darwish, H. (2020). Development, psychometric properties, and pilot norms of the first Arabic indigenous memory test: The Verbal Memory Arabic Test (VMAT). *Journal of Clinical and Experimental Neuropsychology*, *42*(5), 505–515. https://doi.org/10.1080/13803395.2020.1773408