Arabiyat : Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban
Vol. 6 No. 1, June 2019, 71-92
P-ISSN: 2356-153X; E-ISSN: 2442-9473
doi: http://dx.doi.org/10.15408/a.v6i1.10926

ANALYSIS OF TEST ITEM ON THE FINAL TEST SEMESTER EXAM ON ARABIC SUBJECTS

Yogia Prihartini[1], Wahyudi Buska[1], Nur Hasnah[2]
[1]Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi, Indonesia
Jl. Arif Rahman Hakim, Simpang IV Sipin, Telanaipura, Jambi, 36361, Indonesia
[2]Institut Agama Islam Negeri (IAIN) Bukit Tinggi, Indonesia
Jl. Paninjauan, Mandiangin Koto Selayan, Bukittinggi, West Sumatera, 26117, Indonesia
Corresponding E-mail: yogia_prihartini@uinjambi.ac.id

Abstract

This research is aimed; (a) to give a description about Arabic final semester test questions arrangements procedure at the Faculty of Tarbiyah and Teacher Training in UIN Imam Bonjol Padang. (b) to give a description about the quality of the test as seen from its validity, reliability, difficulty, and differentiation capacity. (c) to give a description that the lecturers never did an exam questions analysis of Arabic exam subject at Faculty of Tarbiyah and Teacher Training in UIN Imam Bonjol Padang. This research uses the descriptive method with quantitative and qualitative data. The population within this research is the college students of Tadris major in faculty of Tarbiyah and Teacher Training who follows the final semester exam of 2017-2018 tenet years, especially in Arabic test subject. The sample taken during this research is the college students from the Tadris major, who followed the final semester exam on Arabic subject, by using the purpose sampling technique. The data collecting technique used was documentation, interview, and questionnaires. The results of this research are (1) the types of this tests are achievement test from the teaching objectives aspect, teacher-made test from the arrangement aspect, summative test from the time of implementation aspect, written test from the method of performing aspects, and subjective test from the answer scoring aspect, (2) the quality of Arabic test year 2017/2018: (a) validity sufficient, (b) test reliability is high which is 0,74, (c) the difficulty is easy earning it a 'bad' category, which is 45 numbers of questions or 11,11%, (d) the differentiation capacity is bad, which is 45 numbers of questions, or 22,22%, (3) the reason why the lecturers and question maker team did not held a question analysis are as follows: (a) some of the Arabic lecturers in faculty of Tarbiyah and Teacher Training in UIN Imam Bonjol Padang are not scholars of Arabic education major or not lenient with S-1 and S-2 academic degree, (b) they have zero experience in doing test analysis, (c) there is no order from the authorities which motivates the lecturers in doing test analysis, (d) the lack of understanding about analysis, and (e) there are lecturers who have never received training or seminar about evaluation and analysis.

Keywords: *Analysis, Item Test, Final Test, Arabic*

Introduction

Learning implementation, including learning language, as part of educational implementation, is an effort which preparation and execution includes various parts and steps. Aside from a considerable amount of study and identifications about needs that required to be fulfilled and objectives that need to be achieved, learning implementation also concern the suitable learning material which fits the objectives, aside from the teaching method and technique and proper practice as well. A solid learning implementation includes the coordination of tests to receive various feedbacks about the implemented learning. How the learning process can be measured and can be seen the results. Lecturers must focus their attention on being able to measure students towards their achievement and improvement in learning outcomes.[1]

In normal learning implementation, including language lesson, tests have role and place which clearly connected within it, even becoming an inseparable part[2] of it. In the theory of teaching arrangement and planning, teaching is defined as a process which consists of three inseparable main components. Those components are teaching objective, implementation, and result scoring.[3]

As one of the teaching implementation components, teaching result scoring have a role not less important than teaching objective and implementation. Trough this scoring component, the success or failure of learning implementation can be ascertainable. The assessment is normally done by using a set of instruments ordered and used according to certain procedure so that it can produce information as necessary, and believable. The main instrument mentioned is a test, including language test in language learning assessment.

Due to the reasons above, MuchtarBuchori proposed two educational evaluations special objectives:[4]

1. To find out the learning improvement of the students after they followed, experienced, and acquired education for a certain time span.
2. To find out the effectiveness rate of used educational methods for a certain time span.

Didactically, that assessment will need to have at least five types of roles: (a) as a base giver to assess the students' efforts, (b) to give useful informations, in order to know the positions of each students within their groups, (c) gicing an important meterial to decide the student's status, (d) to give the needed students an orientation to find a solution (e) to give a clue about the progress of the teaching objectives.[5]

---

[1]Arlen R. Gullickson, "Teacher education and teacher-perceived needs in educational measurement and evaluation", *Journal of Educational measurement*, Vol. 23, No. 4, 1986, 347-354.

[2] Ahmad Madkur and Dedi Irwansyah, "Students' perceptions of national examination washback: A case study at MTS Daarul 'Ulya Metro", *Al-Ta'lim Journal*, Vol. 25, No. 4, 2018, 153-162. doi:http://dx.doi.org/10.15548/jt.v25i2.405

[3] Djiwandono, *Tes Bahasa dalam Pengajaran,* (Bandung: ITB Press, 1996), 3.

[4] Mukhtar Buchori, *Teknik-Teknik Evaluasi dalam Pendidikan*, (Bandung: Jemmars, 1980), 6.

[5] Anas Sudijono, *Pengantar Evaluasi Pendidikan*, (Jakarta: Raja Grafindo Persada, 1996), 13.

One of the most important components in learning process that became a deciding factor of the quality of education is the wellness of evaluation devices used by teachers to assess the students learning results. Which means that the way the device was made as an evaluation tool have huge influence in the test's quality. The better an evaluation tool is made and applied it will indicate the quality of the test will be better and tested.[6] The test is one of the most used techniques by psychologists for assessment, evaluation, decision making, and diagnosis.[7] The test is also the most used instrument by the teacher to measure the abilities of his students.

Language testing differs from testing in other content areas because language teachers have more choices to make.[8] According to Kasiram: a good evaluation technique must reflect what it evaluates. As a tool which gives information on the formulation of important decisions in teaching process, test is a vital part which needs to be developed according to a certain criteria. That is why the development, implementation, and use of proper test according to occurring norms, will give many benefits to the success of the education process as whole. A proper test has to fulfill the following characteristics: 1) validity 2) reliability 3) difficulty 4) differentiation capacity.[9]

Final semester test is a summative test which was done by a faculty in order to determine the progress of students in understanding the given material, and then to decide the improvement in the students' rank. Whereas reasoning tests are given as assignments in class at the beginning of the course and students are given a number of points for their completion.[10] Final semester test is also meant to determine the college students' quality in every subjects, especially Arabic. Final semester Arabic test done every year, as a tool to measure the result of one's learning, will have to fulfill certain criteria.[11]

An introductory study has been done atfaculty of Tarbiyah and Teacher Training in UIN Imam Bonjol Padang which became the object of this research. There are several phenomenons happened and is a picture of defect which needs to be fixed. Those phenomenons are: (1) some of the Arabic lecturers atfaculty of Tarbiyah and Teacher Training in UIN Imam Bonjol Padang are not scholars of Arabic

---

[6] Tsai, Huan-Chih, Kwang-Ting Cheng, and Sudipta Bhawmik, "Method and system for improving the test quality for scan-based BIST using a general test application scheme", *U.S. Patent*, No. 6, 694, 466, February 17, 2004, 3.

[7] José Muñiz and Dave Bartram, "Improving international tests and testing", *European Psychologist,* Vol. 12, No. 3, 2007, 206.

[8] James D. Brown and Thom Hudson, "The alternatives in language assessment", *TESOL quarterly*, Vol. 32, No. 4, 1998, 653-675.

[9] Moh. Kasiram, *Teknik Analisa Item Tes Hasil Belajar dan Cara-Cara Menghitung Validity dan Reliability*, (Surabaya: Usaha Nasional, 1984), 12.

[10] Jamie L. Jensen and others, "Teaching to the Test...or Testing to Teach: Exams Requiring Higher Order Thinking Skills Encourage Greater Conceptual Understanding", *Educational Psychology Review*, 2014, 7.

[11] Harjanto, *Perencanaan Pengajaran*, (Jakarta: Rineka Cipta, 2000), 284.

education major. (2) They have zero experience in doing test analysis. (3) There are lecturers who have never received training or seminar about evaluation and analysis.

As implications of those phenomenons, the result is the incompletion of test analysis's problem solving. That's why, the validity, reliability, difficulty level, and differentiation capacity of the questions have never been measured, even though it is a crucial thing to do, so that a defect can be detected and fixed.

The benefit of question analysis is that it also gives a picture of 3 educational components, which is objective; how far is the progress of said objective, the learning process; the accuracy of learning method and technique, and evaluation; the kind and technique of evaluation as well as the arrangement of the questions themselves. A follow up act is to fix those 3 educational components, and it is hoped that teaching implementation will have a maximum appeal and efficiency.

That is why, a research on the final semester test's quality, in this case the Arabic test subject, is crucial in order to improve and develop the quality of Arabic test exam. But so far, an analysis on the quality of final semester exam is rarely, if ever, done. This is what encourages the writer to carry out a research on the quality of Arabic final exam atfaculty of Tarbiyah and Teacher Trainingin UIN Imam Bonjol Padang academic year 2017-2018, from the arrangement procedure to the quality verification.

Based on the background above, we can conclude several problems that need to be addressed and furthermore investigated: (a) How is the Arabic final semester test questions arrangements procedure at Faculty of Tarbiyah and Teacher Training in UIN Imam BonjolPadangcarried out. (b) As seen from its validity, reliability, difficulty, and differentiation capacity, what is the quality of the test. (c) Why the lecturers do never did an exam questions analysis of Arabic exam subject atFaculty of Tarbiyah and Teacher Trainingin UIN Imam Bonjol Padang.

Based on the matter formulation, the objectives of this research are: (a) To give a description about Arabic final semester test questions arrangements procedure at Faculty of Tarbiyah and Teacher Training in UIN Imam Bonjol Padang. (b). To give a description about the quality of the test as seen from its validity, reliability, difficulty, and differentiation capacity. (c). To give a description that the lecturers never did an exam questions analysis of Arabic exam subject at Faculty of Education and Teacher Training in UIN Imam Bonjol Padang.

Method

The feasibility rate of Arabic final test research was a descriptive research which aims to create a systematical and factual description about test's arrangement procedure and the quality from its validity, reliability, difficulty, and differentiation capacity point of view, as well as the reasons why lecturers never done an analysis of the Arabic test. The approach used was the qualitative approach which based its

footing of the opinion of; qualitative design is a procedure which produces descriptive data in the form of written or spoken words about people or observed behavior.[12]

A descriptive method with quantitative and qualitative data which based on qualitative approach was used in this research, since this research aims to produce a description of the test arrangement procedure and to analyze the Arabic subject final semester test's quality in faculty of Education and Teacher Training in UIN Imam Bonjol Padang Academic year 2017-2018 by finding the proof of validity, reliability, difficulty level, and the differentiation capacity. And also to give a description of why the lecturers or the question maker team did not conduct a test analysis. If item-writing guidelines are valid, and non standard items affect test scores, conventional item construction techniques in higher education may be contributing to an underestimation of students' true knowledge, skills, and abilities, as well as negatively impacting student progression.[13]

Data Types And Sources

The data types collected within this research were documents from the interview result. Meanwhile, the data sources were: (a) The profile of the research's object, faculty of Education and Teacher Training in UIN Imam Bonjol Padang. (b) Arabic subject final semester exam at the faculty of Education and Teacher Training in UIN Imam Bonjol Padang question sheets. (c) Arabic subject final semester exam, faculty of Education and Teacher Training in UIN Imam Bonjol Padang students' answer sheets. (d) Arabic subject learning curriculum used by the lecturers. (e) Arabic subject module.

Research Sample And Populations

Research population was a group of people, objects, or things that became the source of the sample, or a collection of sample's classification which were connected with scientific research. There are two types of populations, Finite population, and infinite population. The finite population was one with a limited amount of people in it, while the infinite population was the opposite.[14]

In this research, the writer used the finite population, and proportional random sampling was used as a research sample, by selecting ten college students in every Tadris majors which consist of Math Ed, Physical Ed, Biology Ed, and English Ed, which makes the number of students in this research become 50 students, 29 of which are males, and 21 females. The writer did this in order to make light of the research, and it was considered enough.

---

[12] Lexy J Moleong, *Metodologi Penelitian Kualitatif*, Cetakan VIII, (Bandung: Remaja Rosdakarya Offset, 2000), 3.

[13] David J. Caldwell, and Adam N. Pate, "Effects of question formats on student and item performance", *American journal of pharmaceutical education*, Vol. 77, No. 4, 2013, 71.

[14] David J. Caldwell, and Adam N. Pate, "Effects of question formats on student and item performance", 4.

Data Collection Technique

Data collection techniques used in this research is:

a. Interview

According to Moleong, an interview was done with informal talk using several modules and structured and unstructured interview. An informal talk was done on scientific background. The relationship between interviews with the interviewee with normal circumstances, while the question and answers go normally.[15] This technique was aimed to lecturers' or the question maker team to acquire data on the way to decide and arrange Arabic subject final semester exam at the faculty of Education and Teacher Training in UIN Imam Bonjol Padang academic year 2017-2018. In order for this interview process to go smoothly and able to filter the required data accurately, then an interview module needs to be first arranged in the form of questions related to the creation and arrangement of questions of Arabic subject final semester exam at the faculty of Education and Teacher Training in UIN Imam Bonjol Padang.

b. Documentation

According to Alwasilah, documentation method was a device to collect data in the form of notes, transcript, books, magazines, journal, etc. Which means documentation are written things or movies, which consists of memoir, autobiography, diary, textbook journal, will, working paper, article, newspaper, etc.[16] Documentation of this research was in the form of (a) UIN Imam Bonjol Padang profile, (b) Question sheet of Arabic subject final semester exam, (c) Students' answer sheets of Arabic subject final semester exam at the faculty of Education and Teacher Training in UIN Imam Bonjol Padang (d) Arabic subject learning curriculum used by the lecturers (e) Arabic subject module.

c. Questionnaires

Questionnaires were given to lecturers to acquire data about reasons why they have never done data analysis of Arabic subject final semester exam at the faculty of Education and Teacher Training in UIN Imam Bonjol Padang.

Data Gathering Procedure

In order to gather a representative and confirmative data, the data gathering used following procedure:

a. Preparation; creating an interview module as an instrument which consist of the data learned with all of its sub-aspects.
b. By using interview module as a reference –lecturer as a teacher and question maker– as an informant in this research, according to the agreed schedule. And the acquired data were in the form of a report's note.
c. Collecting test sheets as well as all of the students answer sheets.
d. Creating and arranging questionnaires and spreading them to lecturers and question maker team.

[15] David J. Caldwell and Adam N. Pate, "Effects of question formats on student and item performance", 3.

[16] A. Chaedar Alwasilah, *Pokoknya Kualitatif: Dasar-Dasar Merancang dan Melakukan Penelitian Kualitatif*, (Jakarta: Pustaka Jaya, 2002), 155.

Data Analysis And Tabulation

In this research, data analysis and tabulation was done with following procedures:

a. Validity analysis

A test was said to have a high validity if it performs its functions, or give a proper measurement befitting the means of the measurements, or, in other words, a good test is a test that measures what needs to be measured, and nothing else.[17]

This Arabic final semester test is a learning result test, which means that the analysis used to find out the test's quality in validity, the following validity analysis was done:

1) Analyzing the Arabic subject's curriculum
2) Analyzing every questions and classify them according to category of objectives
3) Comparing the questions' reality with the ideal requirement of the curriculum
4) Deciding the contents' validity by looking at the question reality's proportion shift with its ideal requirement.

b. Reliability Analysis

Reliability was translated to a test which has the ability to produce a constant measurement, even when used upon the same target.[18] Data in the form of answer sheet of Arabic final semester exam year academic 2017-2018 is analyzed to decide the reliability since the Arabic test consists of multiple choices and essay, which was why we use Cronbach Alpha (C. Alpha) formulae.

c. Difficulty Level Analysis

A question's difficulty level was measured by the number between 0,00 to 1,00. Index 0,00 shows that a question was extremely hard, while index 1,00 shows that a number was extremely easy. Which means, the closer a question to number 0 index shows that the question was significantly harder, and the closer it is to number 1, the easier it was.

The interpretation toward difficulty number was:[19]

Less than 0,20 : too hard
0,20 – 0,801 : enough
More than 0,80: too easy

d. Differentiation capacity analysis

To decide the differentiation capacity of each question, the following steps are done:

1) Agglomerating students into two groups, the students with the upper score and students with the lower score
2) Calculating the amount of students' total score on each question in each group. The classifications of the discriminate level and its index span was as follows:[20]

0,50 or more : fine
Between 0,20 and 0,50 : lacking

[17] Saifuddin Azwar, *Reliabilitas dan Validitas*, (Yogyakarta: Pustaka Pelajar, 1997), 6.
[18] Djiwandono, *Tes Bahasa dalam Pengajaran,* 98.
[19] Djiwandono, *Tes Bahasa dalam Pengajaran,* 141.
[20] Djiwandono, *Tes Bahasa dalam Pengajaran,* 144.

0	: none
-(negative)	: negative

Findings

Test validity

Test validity is an essential criteria that should be posessed by every kin of tests, including arabic tests. In order to know the quality of test validity of arabic test among students in arabic language and islamic teaching major in UIN Imam Bonjol Padang year 2017/2018, because this test is a learning result test, the right analysis to conduct is content validity analysis. The steps needed to do this are:

1. Analysing the subject curriculum/arabic module.
2. Identifying the main idea of the subject/arabic module.
3. identifying every questions and classifying them based on their directed category.
4. observing the coherence between questions and the main idea.
5. deciding the content validity by looking at proportion of question reality shift, learning aim, and main idea.

The question maker team have went through a proper procedure, but this test have never been tested and showed to experts which results in the lack of balance of material spreading, and the existence of negative differentiation capabilities. Here are several procedures with their defect and their effect:

a. The aim of this final semester test is to measure the success of college students as a whole, and to decide whether they pass or not.
b. the exam material was derived from the Arabic module *Qowaid Al-'Arabiyah Musyasyarah* كتاب القواعد العربية الميسرة ألفه الدكتور نعمان والدكتور أحمد صفوان.
c. The form of test is an objective test consisting of 40 multiple choices questions and five essay questions, with each multiple question contain 2 points and each essay question contains various points. Subjective test is given five kind of score: four points given to a correct answer, three points given to an almost correct answer, two points given for a half-correct answer, one points given to a false answer, and zero score given to no answer.
d. The domains measured in this test are cognitive, affective, and psychomotor, even though cognitive domain is more emphasized.
e. The difficulty level of this test is fine (this can be seen on test quality analysis as seen from the difficulty level on the oncoming explanation)
f. Question creating has been done properly, but the test have never been showed to experts, which result in inexpediency between the questions and the key answer, and the huge amount of question with negative differentiation capacity.
g. The key answer has been provided, but after thoroughly examined, there are inconsistencies between the key answer with the question. As such, the carefulness of the score checker is required.

Below is the descripiton about Arabic subject's main idea and the test sample used:

Table 1.

NO	TOPIC	MAIN IDEA	NUMBER OF TEST ITEMS	TOTAL	%
1	I	(١) تعارف with basic grammar including ضمير مفرد + عَلَم	1	1	2,22 %
2		(٢) تعارف with basic grammar including ضمير and إشارة مفردة	3	1	2,22 %
3		تقديم الأسرة with basic grammar including ضمير متصل مفرد	2	1	2,22 %
4		في البيت with basic grammar including خبر مقدّم	38	1	2,22 %
5		في الحديقة with basic grammar including نعت	4	1	2,22 %
Total				**5**	**11,11**
6	II	كم الساعة with basic grammar including ساعة.	5, 6	2	4,44 %
7		الذهاب إلى المدرسة with basic grammar including فعل مضارع in جملة فعلية with subject (ضمائر) which have already been taught.	11	1	2,22 %
8		كيف نتوضأ with basic grammar including فاعل ,فعل مضارع and مفعول به.	41	1	2,22 %
9		نتعلم الحساب with basic grammar including eleventh and tenth numbers.	9, 10	2	4,44 %
10		مكتبة المدرسة with basic grammar including أن + فعل مضارع.	7,8	2	4,44 %
Total				**8**	**17,78 %**
11	III	الحفل بذكرى ميلاد الرسول with basic grammar including فعل مضارع and جمع مذكر سالم	12,13, 39	3	6,67 %
12		صوم رمضان with basic grammar including تصريف المضارع.	18	1	2,22 %
13		عيد الفطر with basic grammar including فعل ماض, لم and لا النافية.	19, 20, 43	3	6,67 %

14		برنامج الحفل with basic grammar including تصريف الفعل الماضي.	14, 15, 16, 37	4	8,89 %
15		الشهور القمرية with basic grammar that have already been taught.	21, 22, 42	3	6,67 %
16		خالق العالم with basic grammar including ثلاثي مزيد بحرف.	23, 24	2	4,44 %
17		مناظر القرية with basic grammar including ثلاثي مزيد بثلاثة أحرف and ثلاثي مزيد بحرفين.	25, 26, 34, 35, 44	5	11,11 %
18		الزكاة with basic grammar including فعل الأمر.	27, 28, 45	3	6,67 %
19		الحج with basic grammar including الموصول.	17, 29, 30, 31, 32, 40	6	13,33 %
20		مدرستنا with basic grammar including اسم التفضيل.	33, 36	2	4,44 %
Total				**32**	**71,11 %**
Final amount				**45**	**100 %**

From the table above we can conclude that there are 5 items in topic I's main idea (11,11 %), 8 items in topic II (17,78 %) and 32 items in topic III (71,11%). If we compare them with the requested amount, it will be shows as the table below:

Table 2. The Comparison Between Real Questions and Requested Ideal by The Curriculum

No	Topic	Number of items requested	Percentage	Amount of real questions	Percentage
1	I	4,5	10 %	5	11,11 %
2	II	9	20 %	8	17,78 %
3	III	31,5	70 %	32	71,11 %
Total		45	100 %	45	100 %

We can infer from the table above that the amount of questions have fulfilled the requested ideal by the curriculum, even though there is a miscalculation of topic II due to the number of decimals, so we made adjustments for topic I, turning it into 5 items, topic III 32 items, and the rest is 8 items (topic II). From the above analysis we can say that the final's questions of Arabic subject in UIN Imam Bonjol Padang academic year 2017-2018 have fulfilled content validity.

Test Reliability

The questions in UIN Imam Bonjol Padang's final tests that were analyzed have particular variations in its forms and scoring method. Therefore, these tests have a variation known as *dikotomis*(multiple choices) and *non-dikotomis*(essay). The formula used to find out the level of test reliability of tests that possess both dikotomis and non-dikotomis variation is "*Conbach Alpha formula*", with the following procedure:

1. Counting the variant of each questions using the following formula

$$Sn^2 = \frac{\sum(n^2) - \frac{(\sum n)^2}{N}}{N}$$

Sn^2 = Variant for question number-*n*

$\sum(n^2)$ = Score number-*n*

$(\sum n)^2$ = Total score for each items

N = Number of test participants

$$S1^2 = \frac{46 - \frac{46^2}{50}}{50} = 0{,}07$$

And so on until variant number 45 (for a complete information, refer to attachment 4). The result will be shown as follows:

Table 3. Variant of Each Test Questions

Question	Variant	Question	Variant
1.	0,07	24.	0,22
2.	0,12	25.	0,24
3.	0,19	26.	0,15
4.	0,13	27.	0,25
5.	0,24	28.	0,20
6.	0,13	29.	0,20
7.	0,15	30.	0,22
8.	0,17	31.	0,22
9.	0,23	32.	0,16
10.	0,25	33.	0,12
11.	0,17	34.	0,22
12.	0,22	35.	0,25

13.	0,25	36.	0,25
14.	0,18	37.	0,06
15.	0,09	38.	0,22
16.	0,25	39.	0,20
17.	0,25	40.	0,17
18.	0,24	41.	0,86
19.	0,24	42.	1,85
20.	0,18	43.	1,48
21.	0,24	44.	1,56
22.	0,22	45.	0,26
23.	0,25		

2. Counting the total of all variants from each test questions using the following formula:

$\sum Si^2 = S_1^2 + S_2^2 + \ldots\ldots\ldots\ldots S_{45}^2$

Si^2 = Total variants from all test questions

S_t^2 = *Variant* from all test questions

S_1^2 = Variant from question number 1

S_{45}^2 = Variant from question number 45

Which results in:

$\sum Si^2$ = 0,74 + 0,12 + 0,19 + 0,13 + 0,24 + 0,13 + 0,15 + 0,17 + 0,23 + 0,25 + 0,17 + 0,22 + 0,25 + 0,18 + 0,09 + 0,25 + 0,25 + 0,24 + 0,24 + 0,18 + 0,24 + 0,22 + 0,25 + 0,22 + 0,24 + 0,15 + 0,25 + 0,20 + 0,20 + 0,22 + 0,22 + 0,16 + 0,12 + 0,22 + 0,25 + 0,25 + 0,06 + 0,22 + 0,20 + 0,17 + 0,86 + 1,85 + 1,48 + 1,56 + 0,26

= 13,8188

Total variant from all test questions is: 13,8188 or 13,82 (for 2 decimals after zero).

3. counting the total of all variant from the whole test using the following formula:

$$S_x^2 = \frac{\sum\left(X^2\right) - \left(\frac{\sum X}{N}\right)^2}{N}$$

$$S^2_x = \frac{36793 - \frac{1309^2}{50}}{50} = 50{,}4676 \text{ or } 50{,}47 \text{ (two decimals)}$$

4. Counting coefficient reliability using Cronbach Alpha formula.

$$a = \frac{K}{K-1}\left(1 - \frac{\sum S^2}{\sum S_x^2}\right)$$

$$a = \frac{50}{50-1}\left(1 - \frac{13{,}8188}{50{,}4676}\right) = 0{,}7426$$

Therefore, the test reliability of final Arabic test in UIN Imam Bonjol Padang academic year 2017-2018 for Arabic major is 0,74.

In order to easily interpret the level of test reliability, we can refer to the table below:

Table 4. Reliability Interpretations

Reliability	Category
0,00 – 0,29	Very low
0,30 – 0,49	Low
0,50 – 0,69	Moderate
0,70 – 0,89	High
0,90 – 1,00	Very High

From the result of the counting above, we can infer taht the test's reliability coefficient is between 0,70 – 0,89, and therefore it can be categorized as High.

Difficulty level

The result of analysis for each questions' difficulty level in UIN Imam Bonjol Padang's Arabic major's arabic final test a is as follows:

Table 5. Difficulty Level (P) of Each Test Questions

ITEM NO	P	CATEGORY	ITEM NO	P	CATEGORY
1	0,92	Too easy	24	0,32	Medium
2	0,86	Too easy	25	0,62	Medium
3	0,26	Medium	26	0,82	Too easy
4	0,84	Too easy	27	0,48	Medium
5	0,60	Medium	28	0,28	Medium
6	0,16	Too difficult	29	0,72	Medium
7	0,18	Too difficult	30	0,34	Medium

8	0,22	Medium	31	0,32	Medium
9	0,64	Medium	32	0,20	Medium
10	0,48	Medium	33	0,86	Too easy
11	0,78	Medium	34	0,34	Medium
12	0,34	Medium	35	0,52	Medium
13	0,56	Medium	36	0,52	Medium
14	0,76	Medium	37	0,06	Too difficult
15	0,10	Too difficult	38	0,32	Medium
16	0,56	Medium	39	0,28	Medium
17	0,52	Medium	40	0,22	Medium
18	0,38	Medium	41	0,44	Medium
19	0,58	Medium	42	0,39	Medium
20	0,76	Medium	43	0,26	Medium
21	0,42	Medium	44	0,40	Medium
22	0,32	Medium	45	0,33	Medium
23	0,44	Medium			

In the table above, there are 5 questions (Number: 1, 2, 4, 26 and 33) that fall to the low difficulty category, 36 questions (Nomor: 3, 5, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 32, 34, 35, 36, 38, 39, 40, 41, 42, 43, 44, dan 45) that fall to the medium difficulty category (standard), and 4 questions (Nomor: 6, 7, 15, dan 37) that fall to the high difficulty category.

Test questions that have a medium difficulty filled the majority of the test with 36 items or 80% of total questions. This can be considered a good thing since the test became balanced. The only things that need to be revised are the remaining questions with low difficulty (5 items or 11,11%) and questions with high difficulty (4 items or 8,89%).

Further information will be explained in the following table:

Table 6. Interpretation Precentage of Difficulty Level

NO	CATEGORY	ITEMS	PERCENTAGE
1	Low difficulty	5	11,11 %
2	Medium	36	80,00 %
3	High difficulty	4	8,89 %
Total		45	100 %

Difference Power

Based on criteria concerning difficulty index interpretation and difference power, below we provide a status on each test questions used in Arabic final exam in UIN Imam Bonjol Padang academic year 2017-2018.

Table 7. Difference Power (D) of Each Questions

Item No	D	Category	Item No	D	Category
1	0,18	Lacking	24	-0,24	Negative
2	0,29	Moderate	25	0,47	Moderate
3	0,53	Good	26	0,29	Moderate
4	0,18	Lacking	27	0,41	Moderate
5	0,35	Moderate	28	0,12	Lacking
6	0,24	Moderate	29	0,47	Moderate
7	-0,41	Negative	30	0	None
8	0,18	Lacking	31	0,18	Lacking
9	0,47	Moderate	32	-0,18	Negative
10	0,41	Moderate	33	0,24	Moderate
11	0,24	Moderate	34	0,41	Moderate
12	0,65	Good	35	0,41	Moderate
13	0,76	Good	36	-0,06	Negative
14	-0,35	Negative	37	0	None
15	0,18	Lacking	38	0,06	Lacking
16	0,06	Lacking	39	-0,18	Negative
17	0,71	Good	40	0,06	Lacking
18	0	None	41	0,35	Moderate
19	-0,41	Negative	42	0,56	Good
20	0,35	Moderate	43	0,25	Moderate
21	0,29	Moderate	44	0,28	Moderate
22	0,47	Moderate	45	0,10	Lacking
23	0,47	Moderate			

From the table above, we can see that there are 5 questions with good difference power (11,11%), which is the best level of difference power. 20 questions has moderate difference power (44,44%), and while they didn't make much difference, they are still worth to keep since they can differentiate capable and non-capable students.

Questions that lack difference power are 10 in total (22,22%), they require revision if they are to be used again. Questions that doesn't have any difference power (3 items or 6,67%) need to be replaced. Questions that have negative difference power

(7 items or 15,56%) need to be discarded. As such, there are 25 questions that can be reused for the final exam (55,55%).

Difference power category in the form of table is as follows:

Table 8. Difference Power Table

NO	CATEGORY	ITEMS	PERCENTAGE
1	Good	5 items	11,11 %
2	Moderate	20 items	44,44 %
3	Lacking	10 items	22,22 %
4	None	3 items	6,67 %
5	Negative	7 items	15,56 %
Total		45 items	100 %

From the table above, we can coclude that the quality of this arabic final test is **not very good**, this is due to the amount of items that fall to the lacking category (10 items, 22,22%), none category (3 items, 6,67%), and negative category (7 items, 15,56%). This means that around 15,56% of the questions in UIN Imam Bonjol Padang's arabic final test academic year 2017-2018 need to be discarded since they don't possess sufficients difference power. These exam questions need to be revised if they are to be used again.

Reasons why lecturers never did a test analysis: (a) Some of the Arabic lecturers at Faculty of Education and Teacher Training in UIN Imam BonjolPadangare not scholars of Arabic education major or not lenient with S-1 and S-2 academic degree, (b) They have zero experience in doing test analysis, (c) There is no order from the authorities which motivates the lecturers in doing test analysis, (d) The lack of understanding about analysis, (e)There are lecturers who have never received training or seminar about evaluation and analysis.

DISCUSSION

Tests, Role, and Definition

Etymologically, test came from Latin word testum, which means a device to measure ground. According to Sudijono the word test derived from ancient France tes which means "a plate to set aside precious metals".[21] Inenglish, test mean examination.[22]

In a terminological manner, test is a way to organize an assessment in form of one or several tasks, done by one or several students and then produce a result in

[21] Anas Sudijono, *Pengantar Evaluasi Pendidikan*, 66.

[22] Jhon M Echols dan Hasan Sadily, *Kamus Inggris Indonesia*, (Jakarta: Gramedia, 1995), 584.

accordance of the students' behavior or achievement, which can then be compared with other students' score or standard score which was previously set in.[23]

Muchtar Buchori, in line with Nurkancana and Sumartana, says that test is an experiment done to find out the existence of some or several students' study result. So it can be concluded that test is a series of questions which have to be answered or tasks which have to be done, which become a base for how the tested need to answer the questions or do the tasks. Researchers took the conclusion by comparing it with the set in standard or another tested people.[24]According to Arikunto, test is a series of practice or other devices used to measure skill, knowledge, intelligence, or talent possessed by an individual or group.[25]

In academic terms, according to Sudijono test is a way or procedure used in oreder to measure and assess the educational results, in form of tasks such as questions that need to be answered or duties that need to be done, and from which the students' behavior and achievement can be measured.[26] Meanwhile Sudjana propose the functions of test as an assessment device:[27]

1. A tool to find out whether the instructional objectives have been achieved or not.
2. With this function, the assessment needs to refer to instructional objectives' formulation.
3. As a learning process' feedback. Amendment may be done in instructional, students' learning process, and teachers' teaching strategy aim, etc.
4. A base for arranging the students' learning progress, and then reporting it to their parent. In that report, the student's achievements, score, and his or her ability in learning are written.

Normally, a test has two types of functions: (1) as a device to measure students' abilities, (2) as a device to measure the success of learning program. The first one means to measure the progress achieved by the students after a certain time span. Second one means to measure how far the learning program has been achieved.

Tests can be grouped according to the time it was held, especially its connection with learning language. From those criteria, test can be:

a. Entrance test, which is done before or during a language learning program is started.[28]

---

[23] WayanNurkancana, dan PPN Sumartana, *Evaluasi Pendidikan*, (Surabaya: Usaha Nasional, 1986), 25.

[24] Mukhtar Buchori, *Teknik-Teknik Evaluasi dalam Pendidikan*, 8.

[25] Suharsimi Arikunto, *Dasar-dasar Evaluasi Pendidikan*, (Jakarta: Bumi Aksara, 1993), 29.

[26] Anas Sudijono, *Pengantar Evaluasi Pendidikan*, 67.

[27] Nana Sudjana, *Penilaian Hasil Proses Belajar Mengajar*, Cetakan IX, (Bandung: PT. Remaja Rosdakarya, 2004), 3-4.

[28] Mukhtar Buchori, *Teknik-Teknik Evaluasi dalam Pendidikan*, 19.

b. Formative test, is evaluating how far the children's capacity to form based on the given learning material. And the result can be used by the teachers to improve them whether it is about their learning material or their teaching method. According to Sudjina Formative assessment is a scoring done during the end of learning process in order to see the success rate of said learning process, which orientate on the learning process.[29]

c. Summative test, also known as final semester test, is aimed to measure the success of students and a whole, and also used to take an important decision for the students. Summative test is usually done at the end of a learning program and used to measure the effectiveness of the whole program. The result of this summative test is commonly used to appraise the student, whether they can fulfill the determined material understanding rate.[30]

d. Pre test, which is a test to find out how far the students have mastered the materials given. Pre test is also known as introductory test, which is held in order to find out how far the material that will be given have been matered by the students.[31]

e. Post test, is a test to find out how far the students have mastered the given material. Also known as final test, post test is held to find out whether the important materials which was given to them have been mastered or not.[32]

The test mentioned in this research is Arabic final semester test in at Faculty of Education and Teacher Training in UIN Imam Bonjol Padang academic year ٢٠١٧/٢٠١٨, which was a set of questions in Arabic test subject which must be answered by students, in the form of written objectives.

Criteria of a Proper Test

There are three criteria of a proper test, which is valid, reliable, and practical. According to Djiwandono, there are five criteria of a proper test: (1) validity; it is said valid if it can measure what it is meant to measure, (2) reliability; it can be considered reliable if it can be believed, and produce constant result even with multiple testing with the same subject, (3) objectivity; which is when there is no personal element that influence the scoring result, (4)practicality; the test is said to be practical if it is easy to be done and scored, such as not requiring a lot of devices and giving an advantage to students by offering them the chance to answer the easier questions first, (5) discrimination; the test have the capability to differentiate, which means the result can give information about the students' capabilities.[33]

[29] Nana Sudjana, *Penilaian Hasil Proses Belajar Mengajar*, 5.

[30] Moh. Kasiram, *Teknik Analisa Item TesHasil Belajar Dan Cara-Cara Menghitung Validity dan Reliability*, 22.

[31] Anas Sudijono, *Pengantar Evaluasi Pendidikan*, 69.

[32] Anas Sudijono, *Pengantar Evaluasi Pendidikan,* 70.

[33] Djiwandono,*Tes Bahasa Dalam Pengajaran,* 141-144.

Test Arrangement and Preparation

In order to acquire a high-quality and representative test, the maker need to follow these steps:

a. Defining The Objectives

There are several norms that an important language test needs to go through, which is defining the objectives, and creating the lay-out to arrive at the said objective. Before a language test, several filtering steps and experiments are required to produce a good and representative test.

b. Examination Module

Checking teaching stipulations is important since basically the test materials depend on what the students have learned, except if the objective of said test is only the grouping of students within groups according to their capabilities.

During this step, teacher need to limit the amount of questions based on the weight of the material or skill (test conditions) on the decided subjects, which was taken as a time consideration given within material in common test agenda. After that, teacher try to discipline the test topics according to what happened during experiment, and his or her experience about his or her students.

c. Module Inscription

Modules proposed to students are an important factor which helps the materialization of test's validity and reliability. If students are unable to comprehend these modules, then the validity and reliability cannot be materialized.

d. Amendment

Monitoring the test rate before the amendment is important steps which need to be done in avoid the difficulty after test is done. It is easily possible to remove it before the test begun. That is why, an amendment need to be continuously done after the test creation, and also there also need to be a limitation if the student is asked to write their answer on a question sheet or a separate answer sheet.

e. Test Distribution

A proper test need to have a high measurement degree;validity, reliability, difficulty level, and differentiation capacity.To arrive at this understanding, a test need to go through several experiment and extrapolation steps, and this activity is known as test distribution activity.

The distributed test consists of several topics, which contents go in accordance with students' knowledge.The aim of this implementation is for test experimentation. At every moment, tests need to be experimented in order to check the results statistically. That is what it's meant as a given addition from test believability during the last phase of the test.

f. Test Recollection And Its Final Expulsion

It is important in producing a fine, clear, and non-confusing final test since if there's a mistake in the modules, printing, or editing some of the letters, will affect the

productivity of the students. Such as why the printing needs to be clear, clean, and easy to read.

The word *"language"* has various definitions. Aside from saying that language is words spoken or written, language is also a means of communication for humans. Others defined language as nouns, verbs, sentences, expressions, and many others studies in schools.Language is a system of symbols using sounds that can be pronounciated and heard. Arabic have a special function compared to others.Arabic is not only a language that have a high literary value, but it is also a holy language used to communicate God's will. That is why it contains a value incomprehensible by humans. This is the undisputable fact of Arabic.

Arabic and the Qur'an is two different side of the same coin. Learning Arabic is a must for those who want to learn the Qur'an, and learning the language of the Qur'an means learning Arabic. As such, arabc does not only function as a mean of communication between humans, but also for humans to communicate with their God using prayers.

A happy news for us is during its development, Arabic had been formally decided as one of the world's international language. So it is no surprise that Arabic is heavily encouraged to be taught in educational organizations such as schools, whether it's a state school or a private school. This surely had been calibrated with students' general level of knowledge. In many institutions, Arabic had been a primary foreign language subject aside from English

The results of this research are (1) the types of this tests are achievement test from the teaching objectives aspect, teacher-made test from the arrangement aspect, summative test from the time of implementation aspect, written test from the method of performing aspects, and subjective test from the answer scoring aspect, (2) the quality of Arabic test year 2017/2018: (a) validity sufficient, (b) test reliability is high which is 0,74, (c) the difficulty is easy earning it a 'bad' category, which is 45 numbers of questions or 11,11%, (d) the differentiation capacity is bad, which is 45 numbers of questions, or 22,22%, (3) the reason why the lecturers and question maker team did not held a question analysis are as follows: (a) some of the Arabic lecturers atFaculty of Education and Teacher Trainingin UIN Imam BonjolPadangare not scholars of Arabic education major or not lenient with S-1 and S-2 academic degree, (b) they have zero experience in doing test analysis, (c) there is no order from the authorities which motivates the lecturers in doing test analysis, (d) the lack of understanding about analysis, and (e) there are lecturers who have never received training or seminar about evaluation and analysis.

Conclusion

After the teaching process have been held, a lecturer is required to find out how far the teaching objectives have been fulfilled, or to figure out the capabilities of the college student in understanding the given material. One of the devices to fulfill this means is by using a test. To receive accurate information, a test will need to

possess sufficient criteria, such as validity, reliability, and acceptable differentiation capacity and difficulty level.

The questions analysis of faculty of Education and Teacher Training in UIN Imam Bonjol PadangFinal semester exam is an analysis on summative learning test result made by the members of the question maker team. This research is done to give a description about (1) the types of final semester exam, (2) the quality of test as seen from the validity, reliability, difficulty, and differentiation capacity point of view, (3) the causes of lecturer's lack of understanding towards the test pattern analysis.

The quality Arabic subject final semester exam in UIN Imam Bonjol Padang year academic 2017-2018 is as follows: (a)Test validity is fine, since it fulfill the contain validity standard. (b) Test reliability is high, which is 0,74. (c) From difficulty point of view, 80% of the questions are considered enough, 11% considered too easy, and 8,89% are considered too hard, which makes the test become fine in difficulty category. (d) The differentiation capacity is bad, because of the amount of questions (22,22%) included in lacking category, 6,67% none, and 15,56% negative.

Based on the discussion above it can be concluded that the Arabic lecturers at faculty of Education and Teacher Training in UIN Imam BonjolPadangis fine, since it has already fulfilled the reliability standard and have a high reliability. The test's questions have represented all of the book's material.A test analysis is crucial in order to know whether the test is good or bad. But there are so few lecturers which can be found doing any analysis on test. That is why the writer can take a conclusion that the quality of a test depends on the material comprehension, technique, and question analysis as well as expert and supervisor's analysis.

That is why the writer suggest that the Arabic subject on Faculty of Education and Teacher Training in UIN Imam Bonjol Padang will have to best tested further and require revision or amendment (based on PAP), or removal (based on PAN) especially questions which have no differentiation capability at all. The following is also recommended: 1) improve the test's validity, 2) revise the questions that are too easy, 3) revise the questions which have little to none differentiation capability index, 4) changing the bad questions.[]

REFERENCES

Arikunto, Suharsimi. *Manajemen Penelitian*, Jakarta: Rineka Cipta, 2000.

Azwar, Saifuddin. *Reliabilitas dan Validitas,* Yogyakarta: Pustaka Pelajar, 1997.

Arikunto, Suharsimi. *Dasar-dasar Evaluasi Pendidikan*, Jakarta: BumiAksara, 1993.

Alwasilah, A.Chaedar. *Pokoknya Kualitatif: Dasar-Dasar Merancang dan Melakukan Penelitian Kualitatif*, Jakarta: Pustaka Jaya, 2002.

Brown, James D., and Thom Hudson. "The alternatives in language assessment", *TESOL quarterly,* Vol. 32, No. 4, 1998.

Buchori, Muchtar. *Teknik-Teknik Evaluasi dalam Pendidikan*, Bandung: Jemmars, 1980.

Caldwell, David J., and Adam N. Pate. "Effects of question formats on student and item performance", *American journal of pharmaceutical education,* Vol. 77, No. 4, 2013.

Djiwandono. *Tes Bahasa Dalam Pengajaran,* Bandung: ITB Press, 1996.

Echols, Jhon M.,dan HasanSadily. *KamusInggris Indonesia*, Jakarta: Gramedia, 1995.

Gullickson, Arlen R. "Teacher education and teacher-perceived needs in educational measurement and evaluation", *Journal of Educational measurement,* Vol. 23, No. 4, 1986.

Harjanto. *Perencanaan Pengajaran*, Jakarta: Rineka Cipta, 2000.

Kasiram, Moh. *Teknik Analisa Item Tes Hasil Belajar dan Cara-Cara Menghitung Validity dan Reliability*, Surabaya: Usaha Nasional, 1984.

Madkur, A., & Dedi Irwansyah. "Students' perceptions of national examination washback: A case study at MTS Daarul 'Ulya Metro", *Al-Ta'lim Journal,* Vol. 25, No. 2, 2018. doi:http://dx.doi.org/10.15548/jt.v25i2.405

Moleong, Lexy J. *Metodologi Penelitian Kualitatif*, Cetakan VIII, Bandung: Remaja Rosdakarya Offset, 2000.

Muñiz, José., and Dave Bartram. "Improving international tests and testing", *European Psychologist,* Vol. 12, No. 3, 2007.

Nurkancana, Wayan., dan PPN Sumartana, *Evaluasi Pendidikan*, Surabaya: Usaha Nasional, 1986.

Sudijono, Anas. *Pengantar Evaluasi Pendidikan*, Jakarta: Raja Grafindo Persada, 1996.

Sudjana, Nana. *Penilaian Hasil Proses Belajar Mengajar*, Cetakan IX, Bandung: PT. Remaja Rosdakarya, 2004.

Tsai, Huan-Chih., Kwang-Ting Cheng., and Sudipta Bhawmik. "Method and system for improving the test quality for scan-based BIST using a general test application scheme", *U.S. Patent*, No. 6, 694, 466, February 17, 2004.