Jurnal
Penelitian dan Evaluasi Pendidikan

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 22, No 2, December 2018 (208-218)
Online: http://journal.uny.ac.id/index.php/jpep

HIMPUNAN EVALUASI PENDIDIKAN INDONESIA HEPI

DEVELOPING ASSESSMENT INSTRUMENT OF QIRĀATUL KUTUB AT ISLAMIC BOARDING SCHOOL

Ajeng Wahyuni [1] *, *Badrun Kartowagiran* [1]
[1]Universitas Negeri Yogyakarta
[1]Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
* Corresponding Author. Email: ajengwahyuni77@gmail.com

Abstract

The purpose of this study was to develop and describe the quality and characteristics of an assessment instrument of by name of *Qirāatul Kutub*. This development research is based on Borg & Gall model. The steps of the development were (1) planning, (2) primary developing product, (3) preliminary field testing, (4) main field testing, and (5) final product revision. The subjects for preliminary field testing were 28 testees for performance test instrument and 96 testees for grammar mastery test derived from students of Darul Huda Islamic Boarding School, whereas the subjects for main field testing were 80 students for performance test instrument and 425 students for grammar mastery test. The research instruments were assessment instrument of *Qirāatul Kutub* which consisted of performance assessment instrument and grammar mastery essay test, observation and interview guide. The research data were analyzed based on Rasch Model. The results showed that: (1) the results of the development were 12 items of the performance test and 17 items of the grammar mastery test which were fit with the proposed model, (2) the item difficulty of the performance test ranged from -1.32 to 0.99, while for the grammar mastery test from -1.13 to 1.01; (3) the *person reliability* for the *performance test* was 0.86, *item reliability* was 0.97, and *Cronbach's alpha* value was 0.97. The grammar mastery test has person reliability 0.97 and item reliability 0.99 with Cronbach's alpha 0.90; (4) based on the test information function (TIF) and Standard Error of Measurement (SEM), the instruments were found to be good; the performance test and the grammar mastery test were suitable for learners with ability masteries between -4 and 4.

Keywords: *qirāatul kutub, tarkib, assessment instrument, Islamic boarding school*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v22i2.16592

Introduction

The *Kitab Kuning* (lit.: Yellow Book) is part of a long history of the oldest education institution in Indonesia, the *Pesantren* (Departemen Agama Republik Indonesia, 2003). The *Kitab Kuning*, along with other classical books in Arabic, that has been learned through generations is one of the characteristics of the traditional Islamic boarding school *Salafiyah* (Arifin, 2012, p. 42). The Yellow Book is used as a learning source for various materials such as law, faith, worship, etc. In addition to these purposes, the learning of the Yellow Book requires the mastery of Arabic even in places where Moslems are a minority (Wekke, 2015, p.314). The proficiency in reading the Yellow Book becomes an obligatory skill for students in the traditional *pesantren*. This skill of reading the Yellow Book is called *Qirāatul Kutub,* literary meaning reading the book.

In Arabic learning, *Qirāatul Kutub* is included in the area of language reading skills (*qiraah* = reading). Reading (Nurgiyantoro, 2001, p. 24) is an effort to understand what is told through writing and, thus, knowledge about the writing system, the alphabet, and spelling is important. Meanwhile, Iskandarwassid & Sunendar (2015, p. 246) state that reading is done to get meaning and understanding what is being written. In general, *Qirāatul Kutub* is an activity conducted to understand the meaning and content of the Yellow Book.

As in other reading activities, *Qirāatul Kutub* needs the accompaniment of other skills such as grammar, semantics, and comprehension. Besides, one other important skill of *Qirāatul Kutub* is fluency. These four components then become the criteria for evaluating *Qirāatul Kutub.*

Text reading fluency refers to the quality of reading that is fast and accurate with a natural intonation (Veenendaal, Groen, & Verhoeven, 2015, p. 213). Arabic grammar mainly consists of word forms (Al-Hasyimi, 1971, p. 1). Comprehension is the main component and purpose of reading (Wheeler, Cartwright, & Swords, 2012, p. 416).

In the tradition of the *pesantren*, *Qirāatul Kutub* uses the *Sorogan* method. It is an individual learning method where each learner is faced to the teacher. The learning process runs naturally, proceeding as it does, as it has traditionally and culturally run for hundreds of years. Evaluation is left for the teacher to make to decide whether a student is given a pass or is obliged to re-read the book. This type of evaluation is obviously highly subjective. Learners' achievement is known only to the teacher and decided on only by the teacher.

This subjective evaluation tends to raise questions. According to Nitko & Brookhart (2007, p. 8) evaluation is a process of decision making about the quality and achievement of the learner. To know whether or not an instructional process is successful much depends on the information from an evaluation process. Such information may reveal the quality of the learning product, achievements, and learning difficulties of the students. It may even reveal whether the teacher is successful or not.

The present study is aimed at developing an evaluation instrument for *Qirāatul Kutub* consisting of indicators, assessment criteria, performance test, *tarkib* mastery test, and evaluation report. The presence of a *Qirāatul Kutub* evaluation instrument in the *pesantren* is expected to offer the possibility that evaluation may truly measure success in relation to the objectives and components in the learning. Eventually, actions can be taken as to what aspects can be improved.

Method

The study was developmental research of the Borg & Gall (1989) model with four developmental phases. The first phase was planning which consisted of research and information collecting to include literary reviews related to the problems under study, needs analysis, and preparation for formulating the research framework. In addition, this phase also involved the development of the research procedure, formulation of the objectives to be achieved at each phase, and research design and steps.

The second phase was developing the preliminary form of the product. This phase was intended to develop the initial form of the product. This included preparing for supporting components, guides, and feasibility appraisals of the supporting elements.

The next phase was ppreliminary field testing, which consisted of an initial field testing of a limited scale. The results of this preliminary try-out was used to make improvements of the instrument items. The main field testing was used to reveal the final characteristics of the instrument. The products of the two try-outs were a performance test instrument and a *Tarkib* achievement test instrument of the *Safinah An-Najah* book, both being valid and reliable.

The fourth phase was *operational product revision*, which was revision/perfection of the results of the expanded try-out.

Lastly, the final phase was final product revision. This consisted of the final revision of the developed product.

The product developed in the study consists of several parts. First, the indicators and evaluation criteria are to be used as a guideline by the teacher in giving evaluation to the students' learning. Second, evaluation sheets of *Qirāatul Kutūb* achievement contain essay test items to measure the student's achievement in *tarkib*. And last, evaluation report sheets are used to show the results of the evaluation of the student's performances.

The research subjects were male and female teachers and the students of the Yellow Book learning using the *sorogan* method of the "Darul Huda" *Pesantren* Boarding School in Mayak, Ponorogo. The population consisted of 1,098 students. Respondents for the performance test instrument were five female teachers who were credible in the field, who were members of the examination board for class promotion in *Sorogan*; one as the limited-scale try-out respondent and four as the extended field try0out respondents. The limited try-out involved 28 students for the performance test and 96 students for the *Tarkib* mastery test. The students belonging to the four female teachers were 80 in number. Assignment of the Kretjie & Morgan Table (Wagiran, 2015) for determining the number of the subjects for the essay-type test resulted in 225 students of the marked *Safinah* and 200 of the unmarked *Safinah*. Considering the sample numbers above, sampling was done by simple random sampling within which all students in the group had the equal chance to be selected as a research subject.

In relation to data analyses, qualitative data analyses were used for data derived from the interviews concerning criteria, indicators, and variables. Quantitative data analyses were used for the results of the product try-outs. These analyses were also used to find out the improvements on the product and the validity and reliability measures of the instrument. These analyses also showed the characteristics of the instrument items as well as the achievement of the students.

The first analysis was testing the content validity by expert agreement on the Aiken indexes. According to Retnawati (2016, p. 19), an index is regarded as having a low validity when it is lower than or equal to 0.4, moderate when it is between 0.4 and 0.8, and high when it is higher than 0.8. The second described the quality of the evaluation instrument based on the *Cronbach alpha* value and the Rating Scale Quality Instrument Criteria by looking at the Standard Error of Measurement. Concerning the criteria for the Cronbach alpha value reliability, the instrument will be more reliable as its reliability value is close to 1. Finally, the analyses of test items and evaluation items were done by the Rasch Model since *Partial Credit Model* or PCM was used (Mardapi & Kartowagiran, 2011, p. 329) by the aid of the *Winstep 4.3.* Software. The analyses concentrated on looking at the item fits toward the model (*Goodness of Fit*), level of difficulty, and dispersion of the students' performances.

Levels of difficulty were indicated by item measure values in the logit scale ranging from $-\infty$ to $+\infty$. The higher the logit values, the higher the levels of difficulty (Sumintono & Widhiarso, 2015, p. 70).

The next analysis was conducted on the Test Information Function (TIF) and Standard Error of Measurement (SEM). TIF was used to describe the strengths of a test in conveying the testee's abilities (Retnawati, 2016). It had a reverse correlation with SEM in that the higher the TIF, the lower the SEM. TIF also gave information as how far the test best explained test information. The ability range where the TIF curve crossed the SEM curve was the spot where a test gave the best test information.

The testee's abilities in the analysis output results used the *Winstep 3.73* program in the form of the logit scale. The ability range in the logit scale was from -∞ to +∞. The values of the abilities in the logit scale were then conversed into the 0-to-100 range to make it easy for the teacher or reader to read them. The conversed scores were finally grouped into numbers of categories of ability levels.

Findings and Discussion

The product of the development of the study was a *Qirāatul Kutub* evaluation instrument consisting of 17 essay-type test items of *tarkib* mastery of the *Safinah An-Najah* book. Validation was given by experts in the fields of*Qiraatul Kutub* and assessment, including qualitative item analyses covering contents, construction, and language. The content validity was calculated using the Aiken formula. This phase involved two university lecturers.

Using the Aiken formula for content validity, all the 17 test items for *tarkib* mastery were viable to be used. The Aiken index scores ranged from 0.6 to 1. According to (Retnawati, 2016), an Aiken index score was accepted when it was higher than 0.4.

The results of the field try-out were seen from the test item characteristics and the goodness of fit of the model. This also involved the use of TIF and SEM of the aid of the Winstep 3.73 software. Before the analyses were conducted, three pre-requisite assumptions should be fulfilled of the instrument, namely unidimension, local independence, and parameter invariance. The assumption for unidimension was seen from the Principal Component Analysis (PCA) of the eigenvalue score. When the eigenvalue was higher than 2.0 and the percentages of the variance were higher than the variance percentages of the item, then other factors or dimensions were present and the test was multidimensional. To satisfy the pre-requisite of unidimension, the eigenvalue score should be lower than 2.0 (Linacre, 2011).

From the results of the analyses, it was found that the eigenvalue scores for the first to the last contrasts were lower than 2.0, with a percentage of 5.0%. The conclusion was that there were no other dominant dimensions in the *Qiraatul Kutub* evaluation instrument. The absence of other dominant dimension showed that the instrument measured only one dimension, and thus the unidimension pre-requisite was fulfilled.

The local independence assumption was satisfied when the residual correlation among the items was not higher than 0.3 (Sumintono & Widhiarso, 2015). The results of the analyses showed that none of the items had a residual correlation above 0.3, and thus the pre-requisite for local independence was fulfilled.

The parameter invariance assumption was obtained by dividing the respondents into odd and even groups and comparing their parameters, in the case of the study, levels of difficulty. The results of the analyses showed that the instrument had fulfilled the assumption for parameter invariance.

Reliability of the Instrument

The reliability of the *Qirāatul Kutūb* instrument can be seen in Table 1. The Rasch Model reliability can be seen from the values of the *Cronbach Alpha*, item reliability, and person reliability. It can also be seen, from the separation for item and person, that the higher the separation value, the better the test. Person separation is used to classify testees; if the score is low, it is possible that the instrument is not too sensitive in differentiating between high-scoring and low-scoring testees (Pada, Kartowagiran, & Subali, 2016, p. 9). The reliability criteria are based on the rating scale instrument quality criteria of (Fisher, 2007).

Table 1. Reliability Measure of the *Qirāatul Kutūb* Evaluation Instrument

No	Reliability	Performance Test	Notes	Tarkib Test	Notes
1	Person Reliability	0,86	Good	0,91	Very good
	Person Separation	2,45	Moderate	3,17	Good
2	Item Reliability	0,97	Very good	0,99	Very good
	Item Separation	6,03	Very good	12,98	Very good
3	Test Reliability (Cronbach Alpha)	0,85	Good	0,90	Good

Goodness of Fit

The goodness of fit on the Rasch model was seen from the values of the outfit mean square MNSQ, Z-Standard (ZTSD), and point measure correlation (PR Corr) by the following criteria. The accepted MNSQ value was 0.5 < MNSQ < 1.5, or between 0.5 and 1.5. The accepted ZTSD value was -2.0 < ZTSD < +2.0, or between -2.0 and +2.0. For the Pt measure correlation, the accepted value was 0.4 < Pt Measure Corr < 0.85, or between 0.4 and 0.85. According to Sumintono & Widhiarso (2015, p. 73) an item is considered to fulfill the fit if it has the three criteria. However, an item can still be retain if it has one or two of the three. Results of analyses of the data from the field try-outs are explained through Tables 2 and 3.

Table 2. Goodness of Fit of Field Try-outs Performance Test

Item	Outfit		Pt-Measure	Notes
	MNSQ	ZTSD	Corr.	
1	0,99	00	0,71	Fit
2	0,62	-2,9	0,83	Fit
3	0,76	-1,7	0,63	Fit
4	0,78	-1,5	0,58	Fit
5	0,77	-1,6	0,72	Fit
6	0,96	-0,2	0,70	Fit
7	0,90	-0,6	0,64	Fit
8	1,02	0,2	0,63	Fit
9	1,35	2,2	0,61	Fit
10	0,82	-1,1	0,69	Fit
11	1,37	2,1	0,22	Fit
12	1,44	2,4	0,20	Fit

In Table 2, it can be seen that 12 items of the performance test have a fit although two items, 11 and 12, do not satisfy the criteria for ZTSD and Pt. Corr. The two items still fulfill the criteria for MNSQ; therefore, they may still be retained (Sumintono & Widhiarso, 2015, p. 73).

Table 3. Goodness of Fit of *Tarkib* Mastery Test

Item	Outfit		PT-Measure	Notes
	MNSQ	ZSTD	Corr.	
b1	1,24	3,0	0,63	Fit
b2	0,80	-3,1	0,68	Fit
b3	0,96	-0,4	0,55	Fit
b4	0,91	-1,0	0,62	Fit
b5	0,96	-0,5	0,57	Fit
b6	1,01	0,2	0,62	Fit
b7	1,12	1,7	0,65	Fit
b8	0,97	-0,4	0,61	Fit
b9	1,17	2,2	0,53	Fit
b10	1,04	0,4	0,62	Fit
b11	1,00	0,0	0,62	Fit
b12	1,12	1,6	0,58	Fit
b13	0,92	-1,1	0,67	Fit
b14	0,99	-0,1	0,60	Fit
b15	0,95	-0,5	0,64	Fit
b16	1,07	-1,0	0,63	Fit
b17	0,86	-1,4	0,68	Fit

In Table 3, it can be seen that all the essay-type items of the *tarkib* of the *Qirāatul Kutub* have a fit on the Rasch model, or that the items can be subjected to analyses of this model. As in the above case, although items b1, b2, and b9 do not fulfill the criteria for ZTSD, they satisfy the other two criteria and, hence, can still be retained. Thus, all 17 items have the fit.

Item Characteristics

The item characteristics of the *Qirāatul Kutub* instrument by the Rasch Model was represented by item difficulty levels. Assignment of the categories was done by seeing the Standard Deviation scores. The results are explained through Tables 4 and 5.

In Table 4, it can be seen that items 6 and 7 have the highest difficulty level with a logit value of 1.25. Meanwhile, items 11 and 12 have the lowest difficulty levels with a logit value of -1.76 and -1.95 respectively.

Table 4. Item Difficulty of Performance Test of Field Try-outs

Item	Total Score	Measure (Item Difficulty)	Notes
1	230	-0,80	Moderate
2	209	-0,07	Moderate
3	217	-0,40	Moderate
4	216	-0,37	Moderate
5	186	0,48	Moderate
6	159	1,25	Difficult
7	159	1,25	Difficult
8	165	1,07	Moderate
9	196	0,20	Moderate
10	161	1,19	Difficult
11	258	-1,76	Easy
1212	263	-1,95	Easy

Table 5. Item Difficulty of *Tarkib* Mastery Test

Item	Total Score	Measure (Item Difficulty)	Notes
B1	646	0.19	Moderate
B2	879	-0.32	Moderate
B3	1231	-1.13	Easy
B4	473	0.63	Moderate
B5	1210	-1.07	Easy
B6	792	-0.13	Moderate
b7	862	-0.28	Moderate
b8	727	0.01	Moderate
B9	1005	-0.59	Moderate
B10	383	0.90	Difficult
b11	349	1.01	Difficult
B12	976	-0.53	Moderate
B13	711	0.04	Moderate
B14	1004	-0.59	Moderate
B15	452	0.69	Difficult
b16	639	0.21	Moderate
b17	361	0.97	Difficult

In Table 5, it can be seen that, for the *tarkib* mastery test of *qiraataul kutub* of the *Safinah An-Najah* book, the item difficulty levels range from -1 to 1. Item 11 has the highest difficulty level with 1.01 logit value while item 3 has the lowest difficulty level with a logit value of -1.13.

Information Function (TIF) and Standard Error of Measurement (SEM)

Results of the computation of the data analyses, gave the following curves for information function (TIF) and standard error of measurement (SEM) for the performance test.

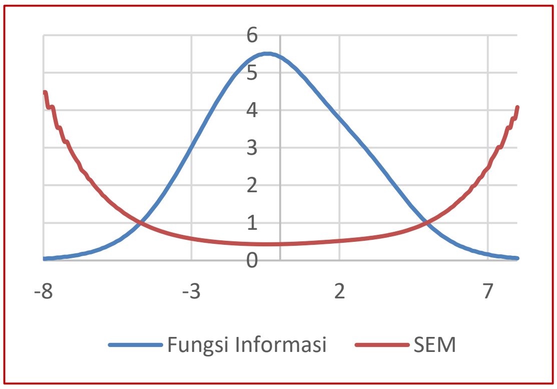

Figure 1. Information Function (TIF) and Standard Error of Measurement (SEM) (Performance Test)

The curves in Figure 1 shows the TIF and SEM of the performance test, wherein the highest is at 5.24 for TIF and 0.24 for SEM. The curves show that the test is suitable for testees with abilities between -4 and 4. The test is said to have good information function as the TIF curve has a reverse form from that of the SEM, in which the TIF highest value crosses the SEM lowest.

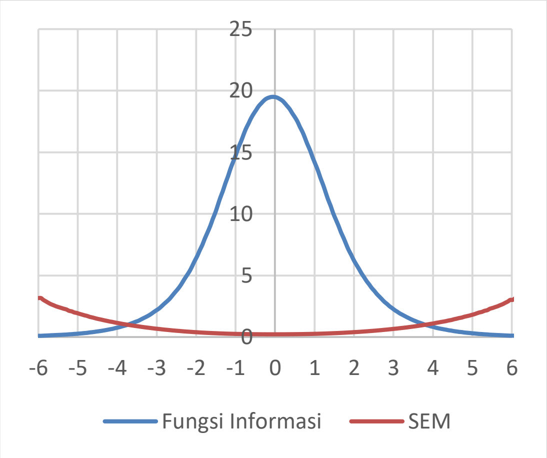

Figure 2. Information Function (TIF) and Standard Error of Measurement (SEM) (*Tarkib* Test)

For the *tarkib* test (Figure 2), the highest TIF is at 19.49 with its SEM 0.11. This means that the test is suitable for testees with moderate abilities between -3.6 and 3.8. Similarly, the test has good information function as the highest value of the TIF is also the lowest of the SEM.

Ability Profile of Safinah An-Najah Book in *Qirāatul Kutūb*

Based on the information from the performance test and *tarkib* mastery test, a general description of the performances of the learners of the Darul Huda boarding school in Mayak, Ponorogo can be explained using the information found in Table 6.

Table 6. Summary of Learners' Mastery (Logit Scale)

No.		Measure (Person)	
		Performance Test	Tarkib Test
1	Mean	-.02	-.23
2	S. D.	1.21	.90
3	Max.	2.31	3.12
4	Min.	-2.81	-2.66

Table 6 shows that the means of the learners' mastery are -0.02 for the performance test and -0.23 for the *tarkib* test. The highest mastery for the *performance test* is 2.31 and for the *tarkib* test 3.12. The lowest logit value for the *performance test* is at -2.81 and for the tarkib test -2.66. Dispersion of the learners' mastery of the performance test can be seen in Figure 3.

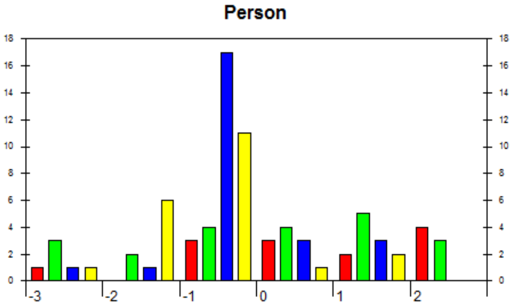

Figure 3. Dispersion of Learners' Mastery (Performance Test)

In Figure 3, it can be seen that the learners' mastery in *Qirāatul Kutub,* based on the Performance Test, stretch from -3 to 3. Most of the abilities lie between logits 0 and -1.

In Figure 4, it can be seen that the majority of the learners have tarkib mastery ranging from -1 to 1, or at the moderate level. Some learners show a mastery of above 2 and the remaining between -1 and -3. The following presents the diagrams for learners' mastery for the performance test in Figure 5 and for the tarkib test in Figure 6.

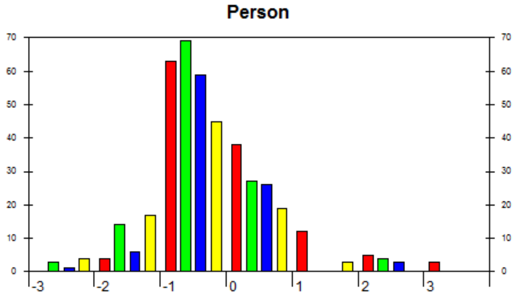

Figure 4. Dispersion of Learners' Mastery (Tarkib Test)

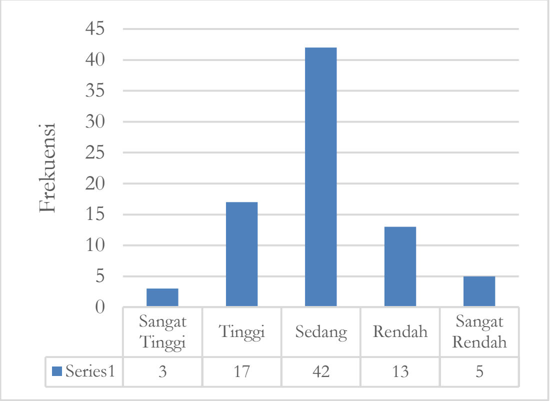

Figure 5. Diagram of Learners' Mastery (Performance Test)

Figure 5 shows that the majority of the learners (42 in number) have mastery of the moderate level. Three learners have mastery of the very high level, and 17 of the high. Then, 5 learners are noted to have a very low mastery level, and 13 others have a low mastery level.

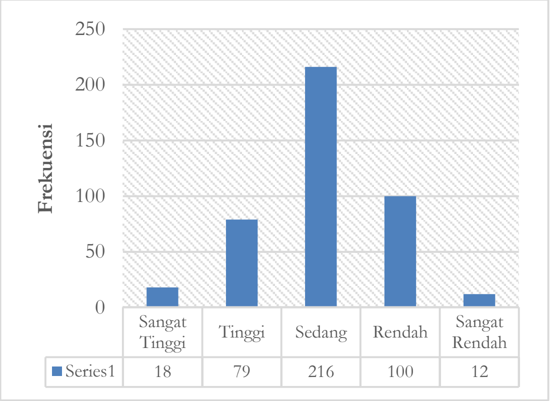

Figure 6. Diagram of Learners' Mastery (Tarkib Mastery Test)

Looking at Figure 6, it can be seen that a large number of learners (256) show a mastery level of the moderate category. The very low category is occupied by 12 learners while the low category 100 learners. On the top levels, 18 learners note the very high category and 79 note the high category. The following presents summary of the learners' mastery in converted scores for both of the tests.

Final Product Revision

Revision of the preliminary draft was given by the validators in the use of vocabulary words from foreign languages (transliteration), such as *mabni*, *tasrif istilahi*, *mubtada' khabar*, and others; and the spelling and writing system of such words. In interrogative sentences, a question must be ended with a question mark. In imperatives, the instruction "sebutkan" (=mention) must be replaced with "tuliskan" (=write down). Another revision was on the answer-key items in the scoring guides. In terms of cla-rity, items 1 and 2 should be merged in order to avoid ambiguities on the part of the respondents.

Table 7. Revision on the Items of the Essay-type Test

Item before Revision	Item after Revision
شهادةانلاالهالااللهوأنّمحمدرسولالله , و**إقام**الصلاةو**إيتاء**الزكاةو**صوم**رمضانو**حج**البيتمناستطاعاليهسبيلا	شهادةانلاالهالااللهوأنّمحمدرسولالله , و**إقام**الصلاةو**إيتاء**الزكاةو**صوم**رمضانو**حج**البيتمناستطاعاليه سبيلا
Item 1: Mention the words that are *mabni* in the paragraph above! Item 2 : Based on item number 1, explain the state of the *mabni* words you have found in the paragraph above!	Item 1 : Write down the words that are mabni in the paragraph above, and give your reasons!

After being revised, and the revision agreed by the experts, the *Qirāatul Kutūb* evaluation instrument was then tried out in the limited forum and then the extended field forum. Following the results of the try-outs, it was found out which items were to be revised and which to be deleted. Based on the results of the data analyses, it was decided that item numbers 11 and 12 could be revised or removed.

Table 8. Revision on Performance Test

Item no	Item description
11	Politeness in behaviour (*Adab*)
12	Politeness in dressing

Field wise, in which *Qirāatul Kutūb* learning was done after the sunset (*maghrib*) prayers, most of the girls were still wearing the *mukena* (attire covering whole body); meanwhile, the learners were sitting leg-crossed in front of the and the teacher was also sitting leg-crossed. In this position, it was hard to conduct evaluation on politeness in both behavior and dressing. As a result, these two items had the lowest level of difficulty and were far different from the other items. In the Goodness of Fit analysis, these two items satisfied only one of the three criteria. This could be understood that these two items were not applicable for a *pesantren* learning condition. However, for possibly a different condition of seating, the items can still be used.

The final product of the *Qirāatul Kutūb* evaluation instrument, there are 12 items for the performance test and 17 for the essay-type.

Final Product

The final product of this developmental research is an evaluation instrument for measuring *tarkib* mastery and a test instrument for measuring performance in *qirāatul kutūb*. The two instruments have undergone two try-outs, first in a limited scope and second in the extended field scope. Beforehand, the drafts of the instruments have also undergone validation processes with experts. The research findings have given indication that the two instruments were valid and reliable and that they have contained items with good quality characteristics. The following presents the validity and reliability of the instruments and the good quality of the item characteristics.

Validity

The instruments are valid since they have fulfilled the criteria for content validity as they have been examined by two experts from the Yogyakarta State University (YSU). In addition, the instruments have been empirically shown to be capable of measuring what they are supposed to measure.

Reliability

The reliability co-efficients of the Cronbach's alpha have been found to have achieved the requirements, 0.85 for the performance test and 0.90 for the tarkib mastery test. The highest information function is at TIF 5.24 and SEM 0.24 for the performance test, suitable for testees with an ability range from -4 to 4. Meanwhile, for the *tarkib* mastery test, the highest value is at TIF 19.49 and SEM 0.11, suitable for testees with an ability range from -4 to 4.

Item Characteristics

Based on the results of the analyses of item characteristics, the two instruments of *Qirāatul Kutūb* have a difficulty of the moderate level with a mean of difficulty levels of 0.00. Summary of the levels of the difficulty measures of the two instruments can be seen in Table 9.

Table 9. Summary of Levels of Difficulty

	Performance Test	Test *Tarkib*
Mean	.00	.00
S. D.	.84	.65
Max.	.99	1.01
Min.	-1.32	-1.13

Table 9 shows the following. For the performance test, the highest difficulty level is 0.99 and the lowest -1.32. Meanwhile, for the *tarkib* test, the highest is 1.01 and the lowest -1.13. The distribution of the difficulty levels of the two instruments ranges from -1 to 1. This shows that the two instruments have a difficulty level of the moderate category.

Conclusion

From the results of the research and development study, the following items of conclusion can be drawn. A number of items can be drawn as conclusion of the study. First, the learning process of *Qirāatul Kutub* runs after the sunset (*maghrib*) prayers, each group consisting of 10 to 20 learners taught by one female teacher. The learning steps run as follows: (1) the teacher opens the session by taking attendance and reciting prayers; (2) one by one, the learners come up to the front of the teacher to submit her reading; (3) the teacher listens to the reading, and asks questions on the *tarkib*, *murad*, or understanding of what has been read; (4) while one learner sits in front of the teacher, the next learners prepare or recite what she is going to submit to the teacher.

The *Qirāatul Kutub* program in the Darul Huda boarding school in Mayak, Ponorogo uses two evaluation formats: a) classroom evaluation conducted by the class teacher, running intuitively without many formal rules and regulations; and b) end-of-year evaluation for class promotion, from which learners' mastery of the Qirāatul *Kutub* is known.

Second, by the Rasch model, quality of the instruments, performance test and mastery test, of the *Safinah An-Najah* is high. The two instruments have passed the validity tests and have empirically been stated as valid. The reliability level of the performance test is marked by a high category of 0.85 and the mastery test 0.90 on the Cronbach's *alpha* co-efficient.. Item and person reliability measures are also found as high. The highest information function TIF is at the 5.24 point with a standard error of measurement SEM 0.24 for the performance test, suitable for learners with ability levels between -4 and 4. Meanwhile, for the mastery test, the highest TIF is at 19.49 with a SEM of 0.11, suitable for learners with ability levels between -4 and 4.

The highest level of difficulty for the performance test is 0.99 and the lowest -1.32. For the *tarkib mastery test,* the highest difficulty level is at the logit value 1.01 and the lowest level -1.13.

Third, the average measure for the learners' mastery is -0.02 for the performance test and -0.23 for the *tarkib* test. The highest learners' performance is at the logit 2.31 and 3.12 for the *tarkib* test. The lowest logit is at the point -2.81 for the performance test and -2.66 for the *tarkib* test. Learners' mastery of the *Qirāatul Kutub* lies between -3 and 3 on the performance instrument. The majority of the learners show mastery between logits 0 and -1. For the tarkib instrument, learners' mastery on the *tarkib* ranges between -1 and 1. A number of learners show a mastery above 2 and the remaining between -1 and -3.

Results of this study is expected to inspire various parties to initiate efforts to develop an instrument applicable for *Qirāatul Kutūb*. For the time being, it is recommendable for teachers and other practitioners of *Qirāatul Kutūb* to use the product of this study, that has gone through careful empirical examinations, for the assessment purposes of *Qirāatul Kutūb* and *tarkib* mastery.

References

Al-Hasyimi, A. (1971). *Qowaidh al-asaasiyah li al-lughah al-'arabiyah*. Beirut: Dar Al-Kotob Al-Ilmiyah.

Arifin, Z. (2012). Perkembangan pesantren di Indonesia. *Pendidikan Agama Islam*, *IX*(1), 40–53.

Borg, W. R., & Gall, M. D. (1989). *Educational research: an introduction* (4th ed.). New York: Longman.

Departemen Agama Republik Indonesia. (2003). *Pondok pesantren dan madrasah diniyah: pertumbuhan dan perkembangannya*. Jakarta: Direktorat Jenderal Kelembagaan Agama Islam.

Fisher, W. P. (2007). Rating scale instrument quality criteria. Retrieved September 20, 2017, from https://www.rasch.org/rmt/rmt211m.htm

Iskandarwassid, I., & Sunendar, D. (2015). *Startegi pembelajaran bahasa*. Bandung: Rosda Karya.

Linacre, J. M. (2011). *User's guide to Winsteps Ministeps Rasch-Model computer programs*. Retrieved from https://www.winsteps.com/winman/copyright.htm

Mardapi, D., & Kartowagiran, B. (2011). Pengembangan instrumen pengukur hasil belajar nirbias dan terskala baku. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *15*(2), 326–341. Retrieved from https://journal.uny.ac.id/index.php/jpep/article/view/1100

Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students*. New Jersey: Pearson Education.

Nurgiyantoro, B. (2001). *Penilaian dalam pengajaran bahasa dan sastra Indonesia*. Yogyakarta: BPFE.

Pada, A. U. T., Kartowagiran, B., & Subali, B. (2016). Separation index and fit items of creative thinking skills assessment. *Research and Evaluation in Education*, *2*(1), 1. https://doi.org/10.21831/reid.v2i1.8260

Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir*. Yogyakarta: Parama Publishing.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan RASCH pada assessment pendidikan*. Cimahi: Trim Komunikata.

Veenendaal, N. J., Groen, M. A., & Verhoeven, L. (2015). What oral text reading fluency can reveal about reading comprehension. *Journal of Research in Reading*, *38*(3), 213–225. https://doi.org/10.1111/1467-9817.12024

Wagiran. (2015). *Metodologi penelitian pendidikan*. Yogyakarta: Budi Utama.

Wekke, I. S. (2015). Antara tradisionalisme dan kemodernan: Pembelajaran bahasa Arab madrasah minoritas muslim Papua Barat. *Tsaqafah*, *11*(2), 313–332.

Wheeler, R., Cartwright, K. B., & Swords, R. (2012). Factoring AAVE into reading assessment and instruction. *Reading Teacher*, *65*(6), 416–425. https://doi.org/10.1002/TRTR.01063