The Development of Higher Order Thinking Skill Test Instrument on the *Fiqh* Subject: The Case of a State Islamic Senior High School in West Bangka Regency

Muh Mawangir
Universitas Islam Negeri Raden Fatah

*Fiqh studies always develop every time. There are various kinds of problems occurring in society regarding the Fiqh of worship and the Fiqh of M*uamalah*, both theoretically and practically. The flexibility of Fiqh studies certainly cannot be addressed with rigid memorization. Of course, it needs understanding, analysis, and synthesis, so that students are ready to respond to the times related to Fiqh. The results of this study showed that the development of the HOTS instrument on Fiqh subject can be carried out by adopting a 4 D development model, consisting of define, design, develop, and disseminate. These results prove that overall, the items have a good level of difficulty since they are in the range of 0.3-0.7. The discriminatory power of the HOTS instrument questions for the Fiqh subject has an average of 0.430. These results prove that the developed HOTS instrument has met the criteria as a good or appropriate item because it has a discrimination index of > 0.3.*

Keywords: designing, developing, disseminating, Fiqh subject, HOTS, state Islamic senior high school, test instrument

INTRODUCTION

The world of education cannot be separated from three essential components, namely teachers, students, and curriculum. Teachers have a strong influence on education (Muazza et al., 2018: Muazza et al., 2019; Mukminin et al., 2019; Sulfemi, 2015). The teacher becomes a facilitator who guides students to move forward. Teachers whose qualified competencies can be good facilitators for students to become educated children. Law No. 14 of 2005 concerning teachers and lecturers explains that teachers are professional educators with the main task of educating, teaching, guiding, directing, training, assessing, and evaluating students in early childhood education through formal, basic education, and secondary education.

Based on Law No. 14 of 2005, one of the duties of a teacher is to evaluate students. Evaluation activities cannot be separated from learning activities. In etymology, evaluation is an assessment (Echols & Shadily, 2000). Meanwhile, in terms of terminology, evaluation is defined as a planned activity to find the condition and the result of an object using instruments compared with benchmarks to obtain conclusions (Yunanda, 2009). Thus, evaluation is an activity of measuring and estimating an object through measuring instruments. An instrument in evaluation activities is essential. If the instrument used in evaluation activities has good characteristics, it will provide valid information.

However, the teacher has paid more attention to the material delivery, so evaluation activities tend to be neglected. It is supported by the interview results with Fiqh subjects teachers at Al-Islam Kemuja State

Islamic Senior High School, Bangka Regency, where the questions made for Fiqh subjects were not from the precise stages. The questions are developed by the teacher and then directly used in measuring students' abilities. The development of questions without precise stages can risk the data collected do not base on actual capability. As a result, the evaluation process of student learning does not produce accurate data. If these results are used to make policies, the resulting policies will not be on target.

Besides, based on interviews and documents, the mid-semester exam and final semester exam for Fiqh subjects constructed by the teachers at all State Islamic Senior High Schools in West Bangka Regency are dominated by questions that measure lower-order thinking skills. It can be seen from the mid-semester and final semester exams in the previous semester that was dominated by questions about memorization. On the other hand, these findings prove that the ability of teachers to make HOTS-based questions is still relatively low. It is also confirmed by the results of research conducted by Retnawati, Djidu, Kartianom, and Anazifa (2018) that stated the ability of teachers to develop HOTS-based questions is still relatively low (Retnawati et al., 2018). Generally, the difficulties faced by teachers in making questions are a lack of understanding of procedures, limited costs, and time. In addition, creativity in realizing questions, especially questions that require HOTS (Suprananto, 2012). Questions about memorization are not able to train the students' reasoning abilities, and it will have an impact on the low level of students' HOTS. In fact, in this disruptive era, the ability to think at a higher level is a crucial asset in competing in the world of work. Wagner (2008) stated that one of the skills that must be owned in the 21st century is the ability to think at a higher level.

The credibility of a test can be seen from its ability to provide a clear description of the program's success level or learning objectives. Some characteristics must be met by an instrument, such as validity, reliability, and practicality (Widoyoko, 2011). Furthermore, to make the goals can be measured, it is necessary to formulate operational learning achievement indicators. However, there are still instruments constructed without considering their characteristics. It needs to be a concern by officeholders, academics, and practitioners to synergize in training and improve the teachers' ability, especially in developing instruments.

Fiqh studies always develop every time. There are various kinds of problems occurring in society regarding the Fiqh of worship and the Fiqh of Muamalah, both theoretically and practically. The flexibility of Fiqh studies certainly cannot be addressed with rigid memorization. Of course, it needs understanding, analysis, and synthesis so that students are ready to respond to the times related to Fiqh. Therefore, the Fiqh subject becomes the research object and development to improve the test instrument quality adapted to the current context and by looking at the various problems that exist in society in general. The author chooses the material of the marriage provisions in Islam as an object that is worth to be developed in this test instrument development. The author considers that it is a material that the development of the questions model can be adjusted in the context of the developed problems in today's society and can provide solutions. Based on the interviews and documents at all State Islamic Senior High Schools in West Bangka Regency, the students' Fiqh learning test results were still dominated by low-order thinking skills, although the 2013 curriculum had encouraged Higher Order Thinking Skills, so the students are less trained to work on questions that measure HOTS. The discrepancy between what should be and what is in the field related to learning outcomes tests makes it needs to have an appropriate and fast solution.

RESEARCH METHODOLOGY

The research method used in the study is Research and Development. This design refers to the 4-D development design. This 4-D development model consisted of four main stages, namely: define, design, develop, and disseminate. The development model was chosen based on the suitability of the research objectives to produce a product in the form of a HOTS test instrument. The product was then tested for feasibility by using a validation process and product testing to determine the feasibility of the test instrument on Fiqh subjects about marriage provisions in Islam based on Higher Order Thinking Skills (HOTS). The stages of development were narrowed down into a 3-D form to the scope of development which only

includes trial of test instruments in a limited, so the design only consists of three stages, namely define, design, and develop.

In broad lines, the main activities in this research are as follows. In pre-development research, it began with a preliminary study through observation and various sources study that was used as a reference to obtain the first information regarding the needs, conditions in the field, and the feasibility of the instruments. It was for designing and developing products. A preliminary study at the pre-development stage summarizes the conclusions of the basics of development to produce product specifications in the form of a test instrument based on Higher Order Thinking Skills on legal material on marriage provisions in Islam along with its wisdom in society at class XI of State Islamic Senior High School. After developing the product, the following step is to make a product validation test design by a test instrument expert and the teacher. This product design validation test was previously validated by a validator (a lecturer and an expert in making test instruments). The product was validated by practitioners (Fiqh subject teachers) before the product validation test was carried out in the field. The validation test results were then analyzed and revised if there were errors. After that, a final product will be produced.

RESULTS AND DISCUSSION

Initial Product Development Results

The procedure of developing an achievement test in learning Fiqh subjects adopts the 4 D (four D) development model proposed by S. Thiagarajan, Dorothy S. Semmel, and Melvyn I. Semmel. The results of the development of each stage are as follows. The instrument grid for the tested Fiqh subjects was formulated based on the curriculum used at Islamic Senior High School Al-Islam Kemuja, which is the 2013 curriculum. Meanwhile, the learning achievement test instrument was constructed based on the main material, namely marriage provisions in Islam. Furthermore, based on this material, the HOTS instrument grid is arranged based on several components such as core competencies, basic competencies, subject matter, indicators, cognitive levels, and answer keys.

Arranging Questions

There were 18 questions of HOTS for Fiqh subjects. Three of them were reserve questions with four alternative answers. The questions were developed by adjusting the cognitive levels of Bloom's taxonomy, consisting of the levels of analyzing (C3), evaluating (C4), and creating (C5). The distribution of the question in terms of cognitive level can be seen in Table 1.

TABLE 1
THE DISTRIBUTION OF QUESTIONS BASED ON THE COGNITIVE LEVEL

Cognitive Level	Items of Questions	Number of Questions
C4 (analyzing)	1,2,3,6	4
C5 (evaluating)	4,5,7,8,9,10,11,12	8
C6 (creating)	13,14,15,16*,17*,18*	6

Table 1 shows that there are 18 questions of HOTS, and three of them are reserve questions (items 16, 17, and 18). Based on Table 1, there are 22% of questions developed at level C4 (analyzing), 33% developed at level C6 (creating), and 45% at level C5 (evaluating). The study of the HOTS questions in the Fiqh subject was carried out based on the assessment of two experts (expert judgment), consisting of Arabic language education experts and education evaluation experts. The questions analyzed used a quantitative approach through the assessment sheet given to the expert. There are several assessment components, which are the suitability of questions with core competencies, basic competencies, cognitive levels, and question indicators. The results of the assessment by the expert were then analyzed using the Aiken's V formula to know the amount of instrument content validity. The results are presented in Table 2.

TABLE 2
THE SUMMARY OF EXPERT REVIEWS

No.	Expert Assessment		Aiken Index
	Education Evaluation	Islamic Religious Education	
1	4	3	0,83
2	3	4	0,83
3	4	3	0,83
4	4	4	1,00
5	4	3	0,83
6	4	4	1,00
7	4	4	1,00
8	4	4	1,00
9	3	4	0,83
10	4	3	0,83
11	4	3	0,83
12	4	3	0,83
13	4	3	0,83
14	3	4	0,83
15	4	4	1,00
16*	3	3	0,67
17*	3	4	0,83
18*	3	4	0,83
Average			0,87

Based on Table 1, the assessments results of two experts indicate that all the HOTS questions in the Fiqh subjects are valid. The validity is indicated by the Aiken index of 0.67 to 1.00. There is 1 question of moderate (number 16), and there are 17 questions with high validity. Meanwhile, overall, the Aiken index of HOTS questions for Fiqh subjects at class XI of State Islamic Senior high School was 0.87. The results prove that the HOTS measuring instrument has high content validity. Retnawati et al. (2018) stated that an index of 0.4 – 0.8 can be classified as media core validity (medium), and an index > 0.8 can be classified as high validity.

Based on the results of a small-scale trial conducted on the 37 students at class XI of State Islamic Senior High School 1 Muntok City, it can obtain results related to empirical instrument validity, reliability, level of difficulty, and discriminatory power of questions. The validity of the questions is empirically known through the product-moment correlation test. Testing is done by connecting the total score with the question score. A summary of the test results using the SPSS 22 program is in Table 3.

TABLE 3
THE SUMMARY OF QUESTION VALIDITY OF THETEST RESULTS

No	r-count	r-table	Interpretation
1	0,373	0,32	Valid
2	0,436	0,32	Valid
3	0,206	0,32	Invalid
4	0,494	0,32	Valid
5	0,403	0,32	Valid
6	0,584	0,32	Valid
7	0,466	0,32	Valid

8	0,650	0,32	Valid
9	0,408	0,32	Valid
10	0,593	0,32	Valid
11	0,460	0,32	Valid
12	0,590	0,32	Valid
13	0,406	0,32	Valid
14	0,473	0,32	Valid
15	0,388	0,32	Valid
16*	0,346	0,32	Valid
17*	0,358	0,32	Valid
18*	0,444	0,32	Valid

Table 3 shows that question number 3 has an r-count/Pearson correlation index of 0.206 and an r-table of 0.32. These results indicate that r-count < r-table, and it means that question number 3 is invalid. The failed question number 3 was not included in the large-scale test because there were reserve questions that had been prepared. Reliability estimation was done using the Cronbach's alpha formula through the SPSS 22 program. The results are in Table 4 below.

TABLE 4
INSTRUMENT RELIABILITY

Cronbach's alpha	N of items
0,772	18

Based on Table 4, the instrument reliability is 0.772, so the conclusion is that the reliability of the HOTS learning outcomes test instrument on Fiqh subjects at class XI of State Islamic Senior High School is high. Retnawati el at. (2018) stated that the closer the alpha coefficient is to 1, the higher the instrument reliability. The level of difficulty of the questions was analyzed using Iteman 3.0 software which was known from the value of Prop. Correct. The test was conducted on 37 students at class XI of State Islamic Senior High School Kemuja who had the same characteristics as Islamic Senior High School 1 Muntok City. The results of the analysis are in the following Table 5.

TABLE 5
THE DIFFICULTY LEVEL OF THE QUESTIONS

Items	The Difficulty Level (*Prop. Correct*)	Interpretation
1	0,676	Enough
2	0,649	Enough
3	0,973	Very Easy
4	0,676	Enough
5	0,676	Enough
6	0,676	Enough
7	0,351	Enough
8	0,568	Enough
9	0,568	Enough
10	0,568	Enough
11	0,405	Enough
12	0,405	Enough

Items	The Difficulty Level (*Prop. Correct*)	Interpretation
13	0,649	Enough
14	0,595	Enough
15	0,676	Enough
16*	0,405	Enough
17*	0,676	Enough
18*	0,541	Enough

Table 5 shows that question number 3 contains 1, which has a very easy difficulty level with an index of 0.973. Based on the results, question number 3 was invalid, so it was not included in the large-scale trial. Allen and Yen (2001) stated that the range of the level of difficulty of a good item is 0.3 – 0.7. Meanwhile, the difficulty level of the other questions is moderate/easy enough because they are in the range of 0.3 – 0.7, so they are maintained and included in the next stage. The discriminatory power of the questions can be seen from the biserial points on the analysis results using Iteman 3.0 software. The results of the analysis of the discriminatory power of the HOTS instrument questions in the Fiqh subject at class XI of State Islamic Senior High School for small-scale trial are in Table 6 below.

Table 6 shows that 1 question has a low discrimination index, 4 questions are moderate, and 10 questions are high. Naga (1992) revealed that the discriminatory power of questions is grouped into three categories, the discrimination index < 0.3 is low, 0.30 – 0.39 is moderate or enough, and > 0.40 is high.

TABLE 6
THE DISCRIMINATORY POWER OF THE QUESTION

Items	Discriminatory Power *(Point biserial)*	Interpretation
1	0,373	Enough
2	0,436	High
3	0,206	Low
4	0,494	High
5	0,403	High
6	0,584	High
7	0,466	High
8	0,650	High
9	0,408	High
10	0,593	High
11	0,460	High
12	0,590	High
13	0,406	High
14	0,473	High
15	0,388	Enough
16*	0,346	Enough
17*	0,358	Enough
18*	0,444	High

Questions with low discriminatory power (<0.3) are invalid and not included in the large-scale trial. It is because questions with low discriminatory power are not able to distinguish between high and low capable students. The large-scale trial is an instrument test at State Islamic Senior High School in West Bangka Regency as the research site. The HOTS instrument was tested on 99 students at class XI, consisting of 19 students of science class (XI IPA), 26 students of social class (XI IPS), and 26 students of MAK 1 XI

class, and 28 students of MAK 2 class XI. The large-scale test results are used as a determinant of the quality of the question from the developed HOTS instrument on Fiqh subjects. The results of this test will also be question parameters stored in the question card that becomes a question bank. The validity of the questions empirically in large-scale trial is known through the product-moment correlation test. Testing is done by connecting the total score with the question score. A summary of the test results using the SPSS 22 program is in Table 7.

Table 7 shows that the 17 questions developed with two questions include the reserve obtained r-count/Pearson correlation index greater than r-table (r-count > r-table). Thus, all questions measuring HOTS in Fiqh subjects are valid. Estimation of instrument reliability in large-scale trial was carried out using the Cronbach's alpha formula through the SPSS 22 program.

TABLE 7
THE RESULT SUMMARY OF THE LARGE-SCALE QUESTIONS TESTING VALIDITY

No	r-count	r-table	Interpretation
1	0,406	0,197	Valid
2	0,406	0,197	Valid
3	0,356	0,197	Valid
4	0,458	0,197	Valid
5	0,528	0,197	Valid
6	0,492	0,197	Valid
7	0,447	0,197	Valid
8	0,447	0,197	Valid
9	0,432	0,197	Valid
10	0,394	0,197	Valid
11	0,407	0,197	Valid
12	0,366	0,197	Valid
13	0,385	0,197	Valid
14	0,388	0,197	Valid
15	0,545	0,197	Valid
16*	0,454	0,197	Valid
17*	0,406	0,197	Valid

The results are in Table 8 below.

TABLE 8
THE RELIABILITY OF LARGE-SCALE TRIAL INSTRUMENT

CRONBACH'S ALPHA	N OF ITEMS
0,725	17

Based on Table 8, the reliability coefficient of the HOTS learning outcome test instrument on the Fiqh subject at class XI of State Islamic Senior High School is 0.725, so the conclusion is that it has met the reliability criteria. Tavakol and Dennick stated that an instrument with a reliability coefficient greater than 0.7 meets the reliability criteria. In addition, Fraenkel, Wallen, and Hyun (2012) also added that the reliability of an instrument should not be less than 0.7.

The Difficulty Level

The difficulty level of the items in the large-scale trial was analyzed using Iteman 3.0 software which was known from the value of Prop. Correct. The results of the analysis are in Table 9 below.

TABLE 9
THE DIFFICULTY LEVEL OF THE QUESTIONS IN LARGE SCALE TRIALS

Items	The Difficulty Level (*Prop. Correct*)	Interpretation
1	0,657	Enough
2	0,677	Enough
3	0,687	Enough
4	0,606	Enough
5	0,566	Enough
6	0,475	Enough
7	0,596	Enough
8	0,596	Enough
9	0,687	Enough
10	0,677	Enough
11	0,394	Enough
12	0,485	Enough
13	0,859	Very Easy
14	0,687	Enough
15	0,687	Enough
16*	0,485	Enough
17*	0,687	Enough

Table 9 shows that the difficulty level of the HOTS instrument questions on Fiqh subject at class XI of State Islamic Senior High School is in the range of 0.3-0.7. There is only 1 question with a difficulty level of > 0.7, which is question number 13 with a 0,859 difficulty index. Thus the question is classified as very easy, so it does not meet one of the criteria for a good question. Allen and Yen (2001) stated that the range of the difficulty level of a good question is 0.3 – 0.7. The analysis of the discriminatory power of questions in large-scale trials was carried out using Iteman 3.0 software. The discriminatory power of the questions can be seen from the biserial point value. The results of the analysis of the discriminatory power of the HOTS instrument items are in Table 10.

TABLE 10
THE DISCRIMINATORY POWER OF LARGE-SCALE TEST ITEMS

Item	Discriminatory Power *(Point biserial)*	Interpretation
1	0,406	High
2	0,406	High
3	0,356	Enough
4	0,458	High
5	0,528	High
6	0,492	High
7	0,447	High
8	0,447	High

Item	Discriminatory Power *(Point biserial)*	Interpretation
9	0,432	High
10	0,394	Enough
11	0,407	High
12	0,366	Enough
13	0,385	Enough
14	0,388	Enough
15	0,545	High
16*	0,454	High
17*	0,406	High

Table 10 shows that all HOTS measuring instrument questions for the Fiqh subject at class XI of State Islamic Senior High School have met the criteria for good discriminatory items, which are in the range >0.30. Naga (1992) revealed that the discriminatory power of questions is grouped into three categories, the discrimination index <0.3 is low, 0.30 – 0.39 is moderate/enough, and >0.40 is high. Based on Table 10, the conclusion is that the developed instrument can distinguish between high and low capable students.

The final product study, the HOTS instrument on Fiqh subject at class XI of State Islamic Senior High School level that has passed the trial and revision/improvement stages, then the final product is formed in the form of a HOTS instrument on Fiqh subjects that have been empirically tested. In today's disruptive era, higher-order thinking skills are an essential asset in competing in the world of work. Wagner (2008) stated that one of the skills that must be owned in the 21st century is the ability to think at a higher level. Therefore, HOTS measuring instrument is needed to know and train students' higher-order thinking skills as developed in this study. The characteristics of the final product questions developed in this study are in the following Table 11.

TABLE 11
THE CHARACTERISTICS OF THE HOTS INSTRUMENT ON FIQH SUBJECT

No	$r_{validity}$	Difficulty Index	Discriminatory Power	Decision
1	0,406	0,657	0,406	Valid
2	0,406	0,677	0,406	Valid
3	0,356	0,687	0,356	Valid
4	0,458	0,606	0,458	Valid
5	0,528	0,566	0,528	Valid
6	0,492	0,475	0,492	Valid
7	0,447	0,596	0,447	Valid
8	0,447	0,596	0,447	Valid
9	0,432	0,687	0,432	Valid
10	0,394	0,677	0,394	Valid
11	0,407	0,394	0,407	Valid
12	0,366	0,485	0,366	Valid
13	0,385	0,859	0,385	Invalid
14	0,388	0,687	0,388	Valid
15	0,545	0,687	0,545	Valid
16*	0,454	0,485	0,454	Valid
17*	0,406	0,687	0,406	Valid
Average	**0,430**	**0,618**	**0,430**	
Reliability		**0,725**		

Table 11 shows that the final product of the development is the HOTS instrument on Fiqh subject at class XI of State Islamic Senior High School, West Bangka Regency, which has met the criteria for good questions. Table 11 also informs that from 17 developed questions, 1 question is invalid (question number 13). Question number 13 does not meet the criteria for the difficulty of a good question because it has a difficulty index of 0.859 that more than the difficulty index range of the question. The good ones are 0.3-0.7. Therefore, question number 13 was replaced with question number 17 as a reserve question, so the final product was a HOTS instrument on Fiqh subjects consisting of 15 questions. Since the product specifications developed at the beginning of the study were 15 questions, these 15 questions meet the criteria as good items from validity, reliability, difficulty level, to discriminatory power and can be used to measure higher-order thinking skills (HOTS) on Fiqh subjects at class XI of State Islamic Senior High School students on the subject of Marriage provisions in Islam, especially at the State Islamic Senior High School in West Bangka Regency.

Fiqh subjects are part of Islamic religious education. Both Fiqh and Islamic religious education are part of education. Etymologically, the word education comes from the Greek language “pedagogy” which consists of two syllables, *pae* means "child" and *ego* means "I guide". Meanwhile, in terms of terminology, education is a conscious effort in developing personality and increasing competence in formal or non-formal education (Arifin, 1876). Hasbullah (2009) defined education as a guidance process consisting of educators, students, goals, and curriculums. Based on the explanation regarding the understanding of education according to the expert, education is a conscious process in developing personality and improving individual competence, which includes educators, students, educational goals, and curriculum.

Government regulation No. 55 of 2007 article 1 explains that religious education is an education that provides knowledge and shapes the students' attitudes, personalities, and skills in practicing their religious teachings, which is carried out at least through subjects/lectures at all paths, levels, and types of education. This government regulation provides opportunities for students to learn about their religious teachings in formal education.

Today, formal educational institutions are developing to suit society's needs by establishing religious schools such as Islamic elementary schools, Islamic junior high schools, Islamic senior high schools, Islamic Boarding Schools, and integrated Islamic schools. Therefore, the applied curriculum related to religious subjects with general subjects tends to be balanced and even more studied, such as Fiqh, Akidah Akhlak, Arabic, and so on, but still within the scope of Islamic Religious Education. Islamic Religious Education is a process of educating students so that they can understand and practice Islamic teachings sourced from the Qur'an and al-Hadith through mentoring, learning, training, and providing experiences. Achmadi (2005) defined Islamic religious education as an effort to develop the nature of students in understanding, living, and practicing the teachings of Islam. Based on expert explanations regarding Islamic religious education, it concludes that Islamic religious education is an effort to guide, nurture, and educate students so that they can understand, appreciate, and practice the teachings of Islam. Meanwhile, Fiqh subjects are part of Islamic religious education. Based on the Regulation of the Minister of Religion of the Republic of Indonesia Number 000912 of 2013 concerning the 2013 Madrasah Curriculum, explains that the Fiqh subject at Islamic Senior High School (Madrasah Aliyah) is one of the Islamic Religious Education subjects which is an improvement from the Fiqh that has been studied previously.

CONCLUSION

Based on the results of the development that has been carried out, The conclusions are as follows. The HOTS instrument development on the Fiqh subject at class XI of State Islamic Senior High School, West Bangka Regency can be done by adopting a 4D development model, consisting of define, design, develop, and disseminate (dissemination). The items of the HOTS instrument in the developed Fiqh subject have met the validity criteria as evidenced by r-count > r-table. The HOTS instrument in the developed Fiqh subjects has met the reliability criteria, as evidenced by the instrument reliability coefficient of 0.725, which is greater than 0.7. The average index of the difficulty level of the HOTS instrument on Fiqh subjects is 0.618. These results prove that overall the items have a good level of difficulty because they are in the range

of 0.3-0.7. The discriminatory power of the HOTS instrument questions on Fiqh subjects has an average of 0.430. These results prove that the developed HOTS instrument has met the criteria as a good/appropriate item because it has a discrimination index > 0.3.

REFERENCES

Achmadi. (2005). *The ideology of Islamic education is the theocentric humanism paradigm*. Yogyakarta: Pustaka Pelajar.

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

Arifin, M. (1976). *The reciprocal relationship of religious education at school with the household*. Jakarta: Bulan Bintang.

Echols, J.M., & Shadily, H. (2000). *Kamus Inggris Indonesia: An English-Indonesian dictionary*. Jakarta: PT. Gramedia Pustaka Utama.

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education*. New York: Mc Graw Hill.

Hasbullah. (2009). *The basics of education*. Jakarta: PT. Raja Grafindo Persada.

Muazza, M., Mukminin, A., Habibi, A., Hidayat, M., & Abidin, A. (2018). Education in Indonesian Islamic boarding schools: Voices on curriculum and radicalism, teacher, and facilities. *Islamic Quarterly*, *62*(4), 507–536.

Muazza, M., Mukminin, A., Rozanna, E.S., Harja, H., Habibi, A., Iqroni, D., . . . Nurulanningsih. (2019). Caring the silenced voices from an islamic boarding school-pesantren: Stories of volunteer teachers and policy implications. *Dirasat: Human and Social Sciences*, *46*(3), 270–279.

Mukminin, A., Habibi, A., Prasojo, L.D., Idi, A., & Hamidah, A. (2019). Curriculum reform in indonesia: Moving from an exclusive to inclusive curriculum. [Kurikularna prenova v Indoneziji: Prehod od izključujočega k vključujočemu kurikulum] *Center for Educational Policy Studies Journal*, *9*(2), 53–72. doi:10.26529/cepsj.543

Naga, D.S. (1992). *Introduction to score theory on educational measurement*. Jakarta: Gunadarma.

Retnawati, H., Djidu, H., Kartianom, A., & Anazifa, R.D. (2018). Teachers' knowledge about higher-order thinking skills and its learning strategy. *Problems of Education in the 21st Century*, *76*(2), 215–230.

Sulfemi, W.B. (2019). The influence of teachers' pedagogic abilities with social studies learning outcomes. *Jurnal Ilmiah Edutecno*, *18*(2), 2302–2825.

Suprananto, K. (2012). *Educational measurement and assessment*. Yogyakarta: Graha Ilmu.

Wagner, T. (2008). *The global achievement gap*. New York. NY: Perseus Books Group.

Widoyoko, E.P. (2011). *Learning program evaluation*. Yogyakarta: Pustaka Pelajar.

Yunanda, M. (2009). *Education evaluation*. Jakarta: Balai Pustaka.