Nadwa: Jurnal Pendidikan Islam Vol. 16, No. 1, 2022
ISSN 1979-1739 (p), 2502-8057 (e)
https://doi.org/ 10.21580/nw.2022.16.1.15384

Developing the Instrument of *Fikih* Assessment in Madrasah Aliyah

Agus Sutiyono[1*]

[1]Walisongo State Islamic University, Indonesia

ARTICLE HISTORY

Submitted
08-03-2022

Accepted
01-04-2022

Published
27-05-2022

ABSTRACT

The study aimed to develop an instrument of *Fikih* assessment in Madrasah Aliyah (MA) that has good construct validity and at estimating the reliability of the *Fikih* assessment instrument in MA. This study used Research and Development (R & D) Method. The subjects in this study were the students in MAN Cilacap, MAN Banyumas, and MAN Purbalingga. This study found, *first*, that the instrument of *Fikih* assessment in the Regency of Cilacap, Banyumas, and Purbalingga has not met its standards as a measurement tool. *Second*, the development of the *Fikih* instrument in the MA has used observable variables by conducting a theoretical review of the *Fikih* learning materials in Grade XI of Senior High School from the 2013 Curriculum. So, the researcher formulated the *Fikih* Instrument Items and conducted a qualitative analysis with the assistance of the *Fikih* experts and Evaluation and Assessment experts. Indicators with 30 items have been equal to 0.976, which means that the nine indicators of the *Fikih* assessment instrument provide reliable estimates for the latent variables of the *Fikih* instrument.

KEYWORDS

instrument development, assessment, *Fikih*

*Corresponding author: Agus Sutiyono ✉(agussutiyono@walisongo.ac.id) Walisongo State Islamic University, Indonesia.

Introduction

Learning process in Madrasah Aliyah is a teaching-learning activity that involves teachers, students, pattern and process of interaction between teachers and students, and learning sources in the learning environment under the fram of educational program implementation. This activity refers to the curriculum that serves as the guideline in conducting the learning process and entails physical and mental activities that both teachers and students perform. Therefore, physical and mental preparation of the teachers in dealing with the students when they deliver the learning materials heavily demands student-active learning. Hasbulah stated that curriculum has been a program, facility, and activity of an educational or a training institution that should be implemented in order to achieve the state objectives (Mukminin et al., 2019) as a program, education is a conscious activiy and is intended to the achievement of curriculum objectives. Assessment for measuring the achievement of the curriculum objectives should be conducted so that the achievement of the objectives can be effective and efficient. The results of the assessment will be feedback and will provide necessary information for implementing the program in the future (feed-forward) (Nurhalim, 2018).

The changes of the students in the domain of knowledge, understanding, skills, and attitudes still demand serious efforts. The changes on those domains have not been maximally pursued and this concern has been perceived by the teachers of *Fikih* in the schools where the study will be conducted. According to the Law of National Education System Number 20 Year 2003, it is mentioned that the government will afford educational quality control through an evaluation system. The law serves as the guideline and the government's role in assuring the improvement of national education quality. As having been written in the Article 57 Verse 1, evaluation is conducted in order to control the educational quality nationally as the form of accountability toward the interested parties.

The manifestation of public accountability in education world is implemented by, for example, performing evaluation as a medium of information. Information that has been gathered from evaluation activities can be: optimization of learning process for achieving learning objectives, teachers' success in conducting the learning process, and students' comprehension capacity in understanding learning materials. However, the three findings are difficult to gather because the research instrument that has been implemented has not been able to perform appropriate measurement. Evaluation basically means performing judgment toward the measurement results; therefore, the errors on assessment and measurement should be made as minimum as possible (Siti Kholidatur Rodiyah, 2019). Educational assessment in Indonesia has been designed on the Government Regulation Number 19 Year 2005 Regarding the Standards of National Education; according to this regulation, one of the most important aspects that should be afforded is the national education standards in relation to mechanism, procedures, and instrument of students' learning results assessment. Similar standards have also

been stipulated in the Ministry of National Education Number 20 Year 2007 Regarding the Standards of Educational Assessment and also in the Ministry of Education and Culture Regulation Number 104 Year 2014 Regarding the Learning Results Assessment by Teachers on the Elementary and High Education.

Based on the information that has been gathered from the pre-survey activities in several Madrasah Aliyahs on the Regency of Cilacap and Banyumas, the researchers have found that the evaluation activities there have not met the standards especially in the process of designing test items. The reality that was found in the sites has shown that the test item for the semester examination has been designed by one of the teachers and will be delivered to all schools. Such condition might cause the test to be less qualified, less valid, and less reliable as a measurement tool. The process of designing test items as a standard tool of learning results measurement becomes one of the educational problems that have been related to the educational quality. As a result, the existing measurement system has not shown good results. Therefore, instrument development becomes highly necessary to be understood and to be implemented by the teachers. In order to develop such instrument, there should be teachers who are productive and and who are able to conduct personal development (Afriana & Festiyed, 2020). In other words, the improvement of educational quality demands the improvement of learning process by implementing systematic procedures and one of the efforts in improving the learning process can be started from improving the assessment system.

A systematic assessment draws the attention of Madrasah Aliyah, which is an educational institution that will deliver the students to the higher education or university level. The statement implies that good graduates can be generated by, for example, habituating the students to be exposed to the good instrument in their assessment so that the students' learning results can be properly assessed. On the other hand, the relationship between the educational quality and the assessment system shows that in the advanced countries the indicators of educational quality are the capacity levels of the graduates that have been generated through a testing system that has been implemented by the authorized institutions. Such relationship indicates that the educational quality can be viewed from the results of external factors testing that has been authorized and the institution that has been entrusted to perform the testing. According to the results of a study by Djemari Mardapi et al., (1999, p.79), assessment system such as National Examination might improve the students' learning motivation, improve the teachers' teaching motivation, locate the school's position, measure the curriculum achievement, identify the improvement of educational quality, and initiate the standardization of educational quality.

Departing from the opinions by several evaluation experts above, the improvement of educational quality can be achieved when there is a good educational evaluation system, for example. Evaluation system such as measurement, assessment, and evaluation in addition to procedure should be systematic; the implementation of

the evaluation system should have high accountability and the results of the implementation are expected to gain recognition from the educational stakeholders. However, what has happened in schools is that the evaluation results will be recognized based on the aspect of evaluation system, which is still problematic. Both the mechanism and the procedure of evaluation system implementation stll suffer from multiple weaknesses. Due to this situation, the evaluation that has been expected to provide input for improving the educational quality is deemed not optimal yet.

Djemari Mardapi et al. (1999, p.79) found several matters that might cause an evaluation system of learning results that has been implemented in the regional schools and the capitol schools to not support the educational quality. These matters are: (1) the quality of the test that the teachers have designed is still less sufficient; (2) the regional testing network has not been well benefitted; (3) the report of test results from the teachers to the principal has not been conducted routinely; and (4) the test results have not been optimally benefitted in order to improve the learning process in the classroom. The regional testing network such as the Association of Indonesian Educational Evaluation (*HEPI, Himpunan Evaluasi Pendidikan Indonesia*) nowadays have developed in neat and accountable manner. This association can be an alternative of reference in designing and developing the assessment instrument. HEPI is an association that has been operating in the domain of evaluation and all aspects that have been related to the problems of measurement instrument. The students' learning achievement will be clearly detected if the instrument that has been implemented has good standards.

However, there are still many teachers who have not implemented an instrument design that is in accordance to the evaluation guidelines. Even there are many teachers who have not understood well the process of designing god test item or instrument. One of the improvements that should be pursued in relation to such problem is designing good instrument for measuring the students' learning results especially for the *Fikih* subject in MA. The results of preliminary survey showed that the mid-semester test item for the *Fikih* subject had been designed by the *Fikih* teachers themselves while the semester test item had been designed by the residency. The residency-level test item designers consisted of 2 *Fikih* teacher from 2 Madrasahs and 1 *Fikih* teacher from another Madrasah who would serve as the editor; as a result, the team would generate a test item that should be implemented throughout the residency. From the information that the researchers had gathered, the test item had not gone through validity and reliability test.

Regarding the use of test form, Djemari Mardapi (2017, p.70) mentioned that the test form that will be implemented consists of objective test and non-objective test. An objective test is able to maintain the objectiveness because the assessor will provide equal score since this type of test has clear scoring guidelines. On the other hand, a non-objective test will be influence by the assessor in terms of scoring system. The two

situations occur because the objective test has objective scoring system while the non-objective scoring system has subjective scoring. Bloom's taxonomy reminds people toward instrument the fact that instrument design should pay attention to the three domains namely cognitive domain, affective domain, and psychomotor domain. In the classification of Bloom's taxonomy, the cognitive domain usually serves as the guidelines by outlining the kind and the type of the test that may be implemented in measuring the students' learning results. A test consists of systematic procedures that aims at observing or describing one or more characteristics of an individual through the use of categorical system or standards. Then, the type of test instrument can be test and non-test. A test is more intended to measure the students' capacity in the cognitive domain. The students' capacity in the cognitive domain might be measured by two methods namely the objective test and the subjective test. Subjective test usually takes the form of essay (explanation); however, in the practice this test cannot cover all of the materials that will be tested. Therefore, in this study the researchers will not implement the subjective test; instead the will implement the objective test. The obective of implementing the objective test is to deal with the weaknesses that have been lest by the essay-form test (Suharsimi Arikunto, 2009, p.162-164).

The use of objective test demands more number of test item in comparison to that of subjective test or essay. According to Suharsimi Arikunto, there are several types of objective test namely: true or false, multiple choice, matching, and completion. Among these types, the researchers will implement the multiple-choice test. A multiple-choice test consists of a statement or a notification about an incomplete understanding. In order to complete the understanding, the students should select one of the available answers. Then the option consists of one key answer and several distractors (Sutarno & Hernawati, 2022).

The advantage of multiple choice test is that this type of test has been able to measure behaviors objectively, whereas the advantage of essay test item is that this type of test is able to measure the students' capacity in organizing their ideas and stating their qanswers according to their own words. On the other hand, the weakness of multiple choice test item is that the distractors are difficult to design (Lions et al., 2021). The understanding toward the test item materials from the basic competencies that have been elaborated into indicators are important in order to adjust the contents of the test item into the document of Basic Competencies and their indicators (Slepkov et al., 2021). A good test instrument should have validity and reliability. The validity of a measurement tool lies on how far the measurement tool has been able to measure what it should measure (Fariyani & Kusuma, 2021; Hasan et al., 2020; Maulida et al., 2020; Setiawan et al., 2021; Nunnally, 1978; Allen & Yen, 1979, p.97; Kerlinger, 1986)). There should be accuracy and appropriate construct in measuring the different students' capacity so that a measuremenr tool might detect the differences among the students. This statement has similar been explained by W. James Popham as follows:

> "It is the validity of a score –based inference that is at issue when measurement folk deal with validity. Test, themselves, do not possess validity. Educational tests are used so that educators can make inferences about a student's status. If high scores lead to one kind of iference; low score tipically lead to an opposite inference. Moreever because validiti hinges on the accuracy of our inferences about students' status with respect to assessment domain, it is technically inaccurate to talk about "the validity of a test." A well-contructed test, if used with the wrong group of examiness or if administered under unsuitable circumstances , can lead to a set of unsound and thoroughly invalid inference. Test based inference may or may not be valid. It is the test-based inference with which educators ought to be concerned (Popham, 1995, p.40-41).

Test validity is an integration of evaluative consideration on the degree of empirical information that has been based on theoretical thought that support the accuracy and the conclusion based on the score test. Prior to administering the test item for measuring and identifying the students' learning results, a qualitative test item analysis might be conducted by means of test review sheet. A very crucial problem in the teacher community, especially the community of *Fikih* teachers in Madrasah Aliyah (MA), is related to the development of an instrument for assessing the students' learning results through the appropriate process. As a result, the already implemented assessment instrument has not been able to provide appropriate information that might serve as feedback toward the programs that have been implemented (feedback). In the same time, the assessment instrument has not also been able to support the future program (feedforward).

Methods

The study was categorized into the Research and Development (R&D). The categorization was based by the opinion of Borg & Gall (1983, p.772) as follows "Educational research and development (R & D) is a process used to develop and validate educational product."

This R&D study made use of several methods namely the descriptive method, the evaluative method, and the survey method. The descriptive method was implemented in order to gather the data regarding the conditions of MA. Then, the evaluative method was implemented in order to evaluate the testingal process within the development of a product. Next, the survey method was conducted in order to describe quantitatively tendencies, behaviors, or opinions of a population by taking the sample of the population. Based on the sample that had been selected, a researcher then would perform generalization (Creswell, 2010, p.216). According to Neuman (2013, p.343), through the survey method a researcher would attain accurate, trustworthy, and valid information or data; however, the survey method should be

implemented under serious efforts and considerations. Within the implementation, the survey method should be conducted by asking a set of questions to a number of people (respondents) regarding their opinions, characteristics, or attitudes in detail.

The study would be conducted in four stages. *First*, the researcher performed a preliminary study namely the activities of gathering information both from the literature and the field review abou the Madrasah Aliyah that had been selected as the research site. *Second*, the researchers designed the draft of the assessment model from the existing dimensions in the Grade XI of Madrasah Aliyah, the guidelines, and the instrument. As having been explained, the design was based on the fact that there had not been any appropriate assessment instrument that might be implemented toward evaluating the students' capacity in the *Fikih* subject. Therefore, the researchers should base the draft of the instrument on the results of their observation. Afterward, the researcher should perform a feasibility test based on the instrument readibility and the expert judgment. *Third*, in conducting the study the researchers performed the limited testing in MAN Purbalingga, MAN Cilacap, and MAN Banyumas by selecting a classroom that consisted of 114 students. Then, the researchers performed the expanded testing in threee MAN by involving 387 respondents. *Fourth*, the researchers revised the final product.

The subjects in the study were the students of Madrasah Aliyah, the teachers of *Fikih*, and the principals of Madrasah Aliyah in the three MAN. In gathering the sample for the study, the researchers had certain considerations regarding the category of the MA. MAN Cilacap belonged to the "Moderate" category, MAN Purbalingga belonged to the "Good" category, and MAN Banyumas belonged to the "Fair" category. The three MAN that had been selected represented the other MA in terms of quality level. The sample was gathered by the researchers through the use of stratified/purposive sampling for each regency. This type of sampling had been the generally implemented method in several the sample under several criteria (Borg & Gall, 1983, p.248).

The total number of the respondents was 387 students, 9 *Fikih* teachers from the three MA, and 3 principals from the three MA. The data gathering methods that the researchers selected in this study were test, documentation, observation, and interview. A test was administered by the researchers in order to gather the data regarding the MA students' comprehensive knowledge. Then, the data that had been gathered as well from the observation, the document, and the interview were the data with regards to the madrasah characteristics such as: achievement, test instrument that had been usually administered, process of designing test instrument, and instrument development.

This research and development study was conducted in two main stages. The first main stage consisted of information gathering activities based on the theory and preliminary product development based on the results of field observation toward the

MA that belonged to the "Good," "Moderate," and "Fair" category. The second main stage consisted of expert judgment and field testing that aimed at identifying the compatibility of the model that had been designed.

In the first main stage or the preliminary stage, the researchers conducted a preliminary study continued by a development study. The activities in the preliminary stage were performed in order to analyse the academic concerns in relation to the madrasah and the teachers' capacity in designing and developing a test instrument. Then, in the second main stage the researchers developed a preliminary product which started from the design of the instrument draft until the instrument testing and analysis. The development of *Fikih* assessment instrument in MA involved the experts and the practitioners of madrasah education. In order to generate a well-qualified instrument, the researchers performed an expert judgment so that the instrument that had been developed could be improved. The validation process was also conducted in order to gain feedback, criticism, and suggestion for the sake of improving the *Fikih* assessment instrument.

After the two main stages had been completed, the researchers performed instrument testing in order to identify the validity and the reliability of the instrument. The testing was conducted in two stages namely the limited testing and the expanded testing. The instrument validity test consisted of several parts. The *first* part was the content validity; through the content validity, the instrument was validated based on the teachers' considerations, the consultation with *Fikih* experts, and the approval of the promotors. The *second* part was the construct validity; through the construct validity, the instrument was validated based on the results of the instrument testing that involved 114 students with the assistance of Iteman program. The limited testing was conducted as a step to define a test instrument that had been validated so that the validated instrument might serve as a measurement tool that met the requirements of test item difficulty level, test item discriminative capacity, and distractor meaningfulness. The test item difficulty level that the researchers had referred to ranged from 0.24 until 0.76. On the other hand, the test item discriminative capacity that the researchers had referred to had several ranges as follows: 0.40 until 1.00 belonged to the "Accepted without Revision" category, 0.30 until 0.39 belonged to the "Accepted with Revision" category, and 0.20 until 0.29 belonged to the "Revised" category. Then, a test instrument which reliability score ranged from 0.60 and 0.70 was possible to belong to the "Accepted" category under the requirement that the instrument had met the indicators of a good construct validity model.

The data analysis technique that had been implemented in the expanded testing was the Confirmatory Factor Analysis (CFA) with Lisrel Program. The confirmatory factor analysis was a model that had been designed under the assumption that describes, explains, or measures empirical data within several relative parameters. This model had been based on the information priority with regards to the data structure in

the form of specific or hypothesized theory (Joreskog & Sorbom, 1993, p.22). In order to test the compatibility between the theoretical model and the empirical data within the assessment instrument, in this study the researchers referred to several criteria of the Goodness of Fit. The RMSEA value that had been less than 0.05 indicated that the model had been fit, while the RMSEA value that had been around 0.08 indicated that the model had reasonable assumption of error. Another criterion which defined that the model had mainly been compatible was the p-value that should be higher than α and the RMSEA should be close to 0.00 (Heri Retnawati, 2016, p.65).

Results

Based on the stages in the research and development study that had been conducted, the researchers would like to discuss the results in several sections below. The first main stage or the preliminary stage consisted of theoretical review and field study which resulted in the selection stage through observation and reading activities. At the end of this stage, the researchers finally selected three Madrasah Aliyah with peculiar conditions in the ex-Residency of Banyumas, namely Cilacap, Banyumas, and Purbalingga. From the observation toward the three MA in this area, the researchers found that the three MA had different quality namely "Good," "Moderate," and "Fair."

Results of Judgment by *Fikih* Expert and Evaluation and Assessment Expert

The researcher consulted the formulation of the guidelines as the material for designing the instrument to the expert of *Fikih* and the consultation involved several experts of Islam namely: 1) Prof. Dr. H. Ahmad Rofiq, MA., from the Science of *Fikih* in Faculty of Syari'ah UIN Walisongo; and 2) Prof. Dr. H. Abdul Hadi, MA., from the Science of *Fikih* in Faculty of Syari'ah UIN Walisongo, 3) Dr. Amir Syamsuddin, M.Ag., a Lecturer of Islam Education from UNY Yogyakarta, 4) Prof. Djemari Mardapi, Ph.D, from Evaluation and Measurement Study Program UNY Yogyakarta, and 5) Prof. Dr. Badrun Kartowagiran, from Evaluation and Measurement Study Program UNY Yogyakarta.

Each indicator would be accepted and might serve as the guidelines in developing the instrument if the expert judgment indicated "Relevant" (R) or "Highly Relevant" (HR) by at least 2 out of 3 *Fikih* experts. The indicators that had fallen into the "Less Relevant" (LR) category would be revised, while the indicators that had fallen into the "Irrelevant" (I) category would be eliminated. In addition, the revision and the elimination of the indicators and/or the aspects was also based on the results of the discussion between the researchers and the experts of *Fikih* and of Evaluation and Measurement who had been mentioned above.

Based on the judgment and the discussions with the three *Fikih* experts who had been involved in the study, there were several indicators that should be revised in terms of language, structure, or heading. On the other hand, there were also several indicators that should be eliminated in order to establish the relevance between the aspects or the indicators and the stages of the Madrasah Aliyah students' cognitive and spiritual development. Other indicators should also be eliminated because these indicators had been overlapping from one to another and had been "inappropriate."

The instrument that had been related to murder should be given special attention because there was a concern that this instrument might influence the students' cognitive aspect. In this instrument, there was a statement which stated that it is appropriate to kill an individual. One of the examples could be found in the following item: "According to the words of the Prophet, we are allowed to kill an individual under three requirements but we are not allowed to kill … ." Then, another example would be as follows: "In the following case of murder, the murderer who should undergo *qishash* is … ." The alternatives of answer to the second example were as follows: a) Johan killed Joni because he thought that Joni had taken away his girlfriend; b) Mr. Usup killed his five-years old son; c) In a brawl among the students, a Seventh Grade student killed his contender; d) When Jiran threw a stone, the stone hit the head of a kindergarten student and the kindergarten student died instantly; and e) Atun died because she was beaten to death by Iqbal using a bottle of mineral water. According to Prof. Dr. Ahmad Rofiq, MA, such items should be given special attention or even should be eliminated. Even at the beginning of his note, he stated that *Fikih* should make use of *syari'i* sentences (certain sentence) instead of ambiguous ones.

Then, Prof. Djemari Mardapi, Ph.D., provided several notes and corrections in composing questions or statements in a test item; the composition should pay attention to the intention of the test item. The composition should be clear so that the students would not have perceptions that might be different from the one intended by the test item. In addition, he also provided notes and corrections for negative questions or statements; the negative questions or statements should be underlined in order to imply clearer meaning so that the students could directly understand the intention of the test item. An example on this matter would be provided as follows: "The followings are not the requirements of a witness in deciding that an individual has committed an affair …" and "The condition in which a husband has promised not to sleep over his wife in certain period is known as …" along with other items.

Prof. Badrun Kartowagiran in his notes and corrections suggested that the scheme that would be consulted to the experts should be clear. For instance, there should be a score of indicator accuracy toward variable and the score of item accuracy toward the indicators. In addition to the notes and the corrections, he also provided several suggestions such as: the statements in a test item should have ambiguous

meanings; a test item should be able to measure the related capacity; and the test item should be functioning well.

Limited Testing

Tabel 1. Limited Testing

Variable	Frequency of Item Acceptance		Total
	Accepted	Denied	
Marriage	10	26	36
Waiting time for Getting Married (*Iddah*)	2	1	3
Inheritance	7	4	11
Murder	4	6	10
Adultery (*Zina*)	7	1	8
Intoxicating Substance (*Khamr*)	3	3	6
Theft	5	1	6
Rebellion (*Bughah*)	2	1	3
Trial	13	4	17
Total	53	47	100

Table 1 showed that from 100 test items there were 53 items that had been accepted and 47 items that had been denied. The items that had been accepted met the requirements of the good test item based on the approach of classical test theory, namely having the difficulty level from 0.30 until 0.80, having discriminative capacity ≥ 0.30, and having effectiveness of functioning distractor (at least 5.00% of the items that had been selected by the participants with low capacity) (Mardapi, 2017, p.129). On the other hand, several items had been denied because these items did not meet the criteria of good test item. Such test items could be too easy or too difficult, had negative and less than 0.30 discriminative capacity, and had ill-functioning distractor. The items that had been accepted would be implemented in the expanded testing (the data gathering process). The results of the expanded testing would be analyzed in order to view the construct validity of the instrument.

Normality Test

Hair (2006) stated that the most fundamental assumption in the multivariate analysis had been the Normality Test, namely a form of data distribution on the single metric variable that resulted in the normal distribution. If the data did not form a normal distribution then the data would be considered abnormal; in other words, the

data would be considered normal if the data formed a normal distribution. The normality could be divided into two parts namely:

1. Univariate Normality
2. Multivariate Normality

The univariate normality was different than the multivariate normality. The univariate normality could be tested by using both the ordinal and the continuous data. On the other hand, the multivariate data could only be tested by using the continuous data. If the data had the multivariate normality, then the data would certainly have the univariate normality. However, if the data had the univariate normality then the data would not necessarily have the multivariate normality.

In order to test whether the assumption of Normality had been violated or not, the researchers would like to use the z-statistic value for the skewness and the kurtosis of the data. The z-value for the skewness could be calculates using the following formula (Ghozali, 2014, p.38).

$$Z_{skewness} = \frac{skewness}{\sqrt{\frac{6}{N}}}$$

```
Relative Multivariate Kurtosis = 1.513

Test of Multivariate Normality for Continuous Variables

             Skewness                      Kurtosis           Skewness and Kurtosis

      Value  Z-Score P-Value     Value  Z-Score P-Value      Chi-Square P-Value
     ------  ------- -------   -------  ------- -------      ---------- -------
    985.787  200.259   0.000  1452.902   30.555   0.000       41037.096   0.000
```

Model Compatibility Test

Hair (2006) stated that the most fundamental assumption in the multivariate analysis had been the Normality Test, namely a form of data distribution on the single metric variable that resulted in the normal distribution. If the data did not form a normal distribution then the data would be considered abnormal; in other words, the data would be considered normal if the data formed a normal distribution. The normality could be divided into two parts namely:

The model compatibility test was intended to assess the accuracy of a model. There were many indicators that had been able to be assigned in assessing a model. However, in this study the researchers implemented the criteria of model fitness based on the value of Root Mean Square Error of Approximation (RMSEA). RMSEA had been the most informative indicator of model fitness. Browne & Cudeck (1993) stated that

RMSEA had been implemented in measuring the deviation of parameter values in a model with the covariance matrix of the population (Mardapi, 1999). The value of RMSEA that had been lower than 0.05 indicated that the model had been fit, while the value of RMSEA that had been ranging around 0.08 indicated that the model had possessed reasonable assumption of errors (Mardapi, 1999). However, the most important matter in stating that a model had been compatible was that the p-value had been higher than α and the RMSEA value had been closer to N referred to the sample size. The z-statistic score for the kurtosis of N could be calculated by using the following formula (Mardapi, 1999)

If the z-scores, both the $z_{kurtosis}$ and/or the $z_{skewness}$, were significant (p-value < 0.05) then the data distribution would be abnormal. On the other hand, if the z-scores, both the $z_{kurtosis}$ and/or the $z_{skewness}$, were significant (p-value > 0.05) then the data distribution would be normal. If the data distribution was abnormal then one of the alternatives that the researchers could pursue was estimating the model based on the Maximum Likelihood and performing correction toward the deviation upon the violation of Normality by using the asymptotic covariance matrix. 0 (Heri Retnawati, 2015, p.65). The results of the model compatibility test could be viewed in Figure 1 as follows.

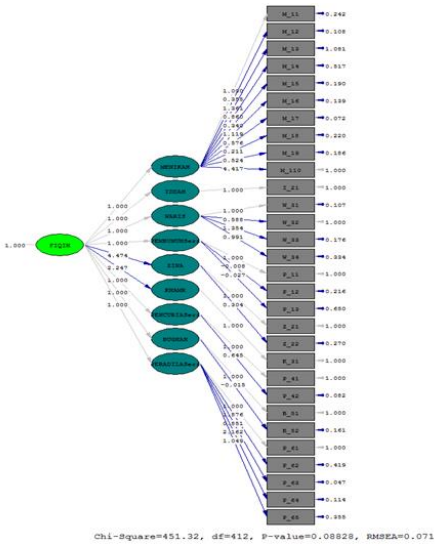

Figure 1. Results of Confirmatory Factor Analysis

Figure 1 showed the results of model compatibility analysis to the p-value that had been higher than α, namely 0.08828 > 0.05000, and with the RMSEA value that had been lower than 0.08, namely 0.07. The findings here might imply that the model had been compatible or the empirical data had been identical to the theory or the model.

Discussion

Construct validity, is the concept of measuring validity by testing whether an instrument measures the construct as expected to developing the instrument of *Fikih* Assessment.

The proof of construct validity could be viewed from the value of convergent validity (loading factor value) on the Standardized Solution. Since the loading factor value > 0.70, then the model met the requirements of good convergent validity. However, according to Retnawati (2016, p.64) the path coefficient would be meaningful if the path coefficient had not been lower than 0.40 and the T-value had not been lower than 1.96 (non-red path). The results of Standardized Solution could be viewed in Figure 2 as follows.

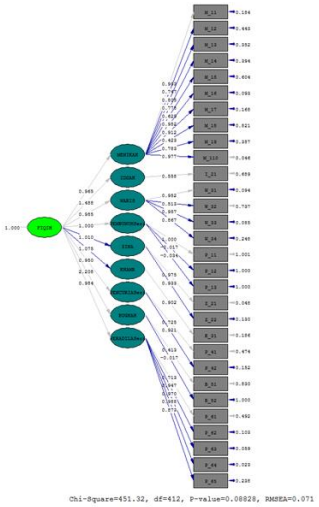

Figure 2. Standardized Solution

Figure 2 showed that all of the observable variables or all items had provided significant contribution to the latent variables (indicators) except the indicator P_12, P_13, and B_52. This finding was shown by the loading factor score (path coefficient) of all observable variables on the latent variables had been higher than 0.40 except the observable variable P_12, P_13, and B_52; the loading factor scores of the three observable variables were respectively as follows: -0.017, -0.034, and -0.017. This finding might imply that the indicator P_12, P_13, and B_52 had not provided meaningful contribution to the latent variable of murder and *Bughah*; as a result, the three indicators should be dropped.

Similarly, based on the T-value, all of the significant path coefficient, except the indicator which T-value had been lower than 1.96, fell into the red line path. On the other hand, in a more in-depth view, the observable variable of Marriage, *Iddah*, Inheritance, Murder, *Zina, Khamr,* Theft, *Bughah*, and Trial had been the dimension of

the observable variable of *Fikih*. This matter was shown by the loading factor value on all factors that had been higher than 0.40 and the T-value that had been 1.96 (black path). The T-value scores were provided in Figure 3 as follows.

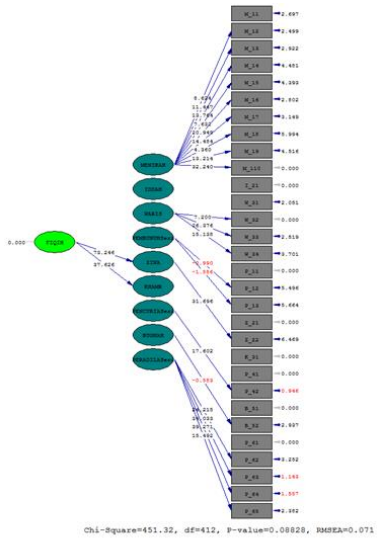

Figure 3. T-Value

Figure 3 showed that several paths did not appear. The reason was that the path from Marriage to M_11, Iddah to I_21, Inheritance to W_31, Murder to P_11, Zina to Z_21, *Khamr* to K_31, Theft to P_41, *Bughah* to B_51, and Trial to P_61 had been fixed so that these paths did not appear. Similarly, the path from *Fikih* to Marriage, Waiting Time for Getting Married (*Iddah*), Inheritance, Murder, Adultery (*Zina*), Intoxicating Substance *(Khamr)*, Theft, Rebellion (*Bughah*), and Trial had been fixed. The paths that had been fixed were made in order that the estimated parameters would be lower and the degree of freedom would increase. The number of degree of that increased would influence the model specification and the model fitness.

Construct Reliability

For testing the construct reliability, the researchers made use of the loading factor from each indicator that composed the instrument (λ) and the index of unique error from each indicator (δ). The formula for testing the construct reliability was as follows (Geldhof, Preacher, & Zyphur, 2014).

$$CR = \frac{\left(\sum_{i=1}^{i} \lambda_i\right)^2}{\left(\sum_{i=1}^{i} \lambda_i\right)^2 + \left(\sum_{i=1}^{i} \delta\right)}$$

Bagozzi & Yi (1988) stated that the cut-off level that could imply that the construct reliability had been good had been 0.60. The results of construct reliability calculation were provided in Table 2 as follows.

Tabel 2. Construct Reliability

Indicator	Λ	δ
M_11	0.903	0.184
M_12	0.747	0.443
M_13	0.805	0.352
M_14	0.778	0.394
M_15	0.629	0.604
M_16	0.952	0.093
M_17	0.912	0.168
M_18	0.423	0.821
M_19	0.783	0.387
M_110	0.977	0.046
I_21	0.558	0.689
W_31	0.952	0.094
W_32	0.513	0.737
W_33	0.957	0.085
W_34	0.867	0.248
P_11	1.000	1.001
P_12	-0.017	1.000
P_13	-0.034	1.000
Z_21	0.975	0.048
Z_22	0.933	0.13
K_31	0.902	0.186
P_41	0.725	0.474
P_42	0.921	0.152
B_51	0.413	0.830
B_52	-0.017	1.000
P_61	0.713	0.492
P_62	0.947	0.103
P_63	0.970	0.059
P_64	0.988	0.023
P_65	0.873	0.236
Total (∑)	22.048	12.079
Square	486.114	145.902
Construct Reliability	0.976	

Table 2 showed that the construct reliability of 9 indicators in item number 30 had been equal to 0.976. This construct reliability might imply that the 9 indicators of *Fikih* provided reliable estimate for the latent variables of *Fikih*. However, there were 3 variables that did not provided meaningful contribution namely the variable P_12, P_13, and B_52. P referred to the variable Murder, while B referred to the variable Rebellion (*Bughah*). In line with the notes by Prof. Ahmad Rofiq, M.A. that test items that had been related to the murder should be given special attention, there was evidence that the empirical data showed that the variable Murder had not enforced the learning materials of *Fikih*. There was also other evidence that the variable Rebellion (*Bughah*) that had been related to dissidence had a closer position to murder and, therefore, this variable had not enforced the learning materials of *Fikih*.

Conclusion

The instrument of learning results assessment for the *Fikih* subject in Madrasah Aliyah on the Regency of Cilacap, the Regency of Banyumas, and the Regency of Purbalingga has not met the standards of a measurement tool. The reason is that the Madrasah Aliyah in the three regencies has limited human resources, especially the *Fikih* teachers, within the instrument development.

The development of an assessment instrument for the learning results of *Fikih* subject in the Madrasah Aliyah has been based on the observable variables and the development has been conducted by performing a theoretical review toward the learning materials of *Fikih* for Grade XI of Madrasah Aliyah under the 2013 Curriculum and a qualitative analysis that involved the expert of *Fikih* and the expert of evaluation and measurement. The quantitative analysis has been conducted based on the data of the empirical testing and the quantitative analysis has been the basis of redesigning the instrument after the instrument has been revised. The results of model fitness test have indicated that the model has been fit into the data. The evidence of the model fitness can be found in the results of model fitness analysis that shows that the p-value has been greater than the α, namely 0.08828 > 0.0500, and that the Root Mean Square Error of Approximation (RMSEA) has been smaller than 0.080, namely 0.071 > 0.080. The RMSEA value that has been around 0.080 indicates that the model has had a reasonable assumption of error. In other words, the model has been fit or the empirical data have been identical to the theory or the model.

The estimation of reliability for the assessment instrument of *Fikih* learning results in Madrasah Aliyah has been equal to 0.976 from 9 indicators with 30 items. The estimation value has indicated that the 9 indicators have provided reliable estimation for the latent variables of *Fikih*.

References

Afriana, I., & Festiyed. (2020). Meta-analysis of authentic assessment instrument development to measure learning outcomes of learners SMA/MA. *Journal of Physics: Conference Series,* 1481(1), 012052. https://doi.org/10.1088/1742-6596/1481/1/012052.

Allen, M. J., & Yen, W.M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole Publishing Company.

Anderson, L., & Krathwohl, D. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Arikunto, S. (2004). *Evaluasi Program Pendidikan*. Jakarta: Bumi Aksara.

Arikunto, S. (2009). *Prosedur penelitian suatu pendekatan praktek*. Jakarta: Rineka Cipta.

Borg, W. R.A., & Gall, M. D. (1983). *Educational research an introduction.4th Edition.* New York: Longman Inc.

Creswell, J. W. (2008). *Educational research; planning, conducting, and evaluating quantitative and qualitative research.* New Jersey: Pearson Educational Edition.

Depdiknas. (2003). *Undang-Undang RI Nomor 20, Tahun 2003, tentang Sistem Pendidikan Nasional.*

Depdiknas. (2005). *Peraturan Pemerintah RI Nomor 19, Tahun 2005, tentang Standar Nasional Pendidikan.*

Depdiknas*. (2006). Peraturan Menteri Pendidikan Nasional RI Nomor 22, Tahun 2006, tentang Standar isi.*

Fariyani, Q., & Kusuma, H. H. (2021). Development of Test Instruments to Analyze Higher-Order Thinking Skills Through Science-Based Literacy Learning. *JIPF (Jurnal Ilmu Pendidikan Fisika),* 6(1), 76. https://doi.org/10.26737/jipf.v6i1.1886.

Geldhof, G.J., Preacher, K., Zyphur, M.J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Physocological methods,* 19 (1).

Gronlund, N.E. (1976). *Measurement and evaluation in teaching.* New York: Macmillan.

Hair, J. F. JR., Anderson, R.E., Tatham, R.L., & Black, W.C. (2006). *Multivariate data analysis*. Six Edition. New Jersey: Pearson Educational, Inc.

Hasan, S. W., Auliah, A., & Herawati, N. (2020). Pengembangan Instrumen Penilaian Kemampuan Berpikir Kritis Siswa SMA. *Chemistry Education Review (CER)*. https://doi.org/10.26858/cer.v3i2.13769.

Joreskog, K., & Sorbum, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software International Inc.

Kerlinger, F. N. (1986). *Asas-asas penelitian Behavioral* (terj. L.R. Simatupang). Yogyakarta: Gajah mada University Press.

Levin, H. M. (2011). *The utility and need for incorporating non-cognitif skill into large scale educational assessment.* Presented at ETS Invitational on International Large-Scale Assessment and to be published by Educational Testing Service. USA: Teacher College, Columbia University.

Lions, S., Monsalve, C., Dartnell, P., Godoy, M. I., Córdova, N., Jiménez, D., Blanco, M. P., Ortega, G., & Lemarié, J. (2021). The Position of Distractors in Multiple-Choice Test Items: The Strongest Precede the Weakest. Frontiers in Education. https://doi.org/10.3389/feduc.2021.731763.

Mardapi, D. (1988). *Practical implementation of validity generalization whit the Indonesian University selection test (sipenmaru).* Disertasi doktor 1988.

Mardapi, D. (2017). *Pengukuran, penilaian dan Evaluasi Pendidikan*. Yogyakarta: Parama Publishing.

Mardapi, D., dkk. (1999). *Survey kegiatan guru dalam melakukan penilaian di kelas.* Laporan Penelitian. Yogyakarta: IKIP Yogyakara.

Maulida, I., Dibia, I. K., & Astawan, I. G. (2020). The Development of Social Attitude Assessment Instrument and Social Studies Learning Outcomes Grade IV on Theme of Indahnya Keragaman di Negeriku. *Indonesian Journal Of Educational Research and Review*, 3(1), 12. https://doi.org/10.23887/ijerr.v3i2.25823.

Mukminin, A., Habibi, A., Prasojo, L. D., Idi, A., & Hamidah, A. (2019). Curriculum reform in indonesia: Moving from an exclusive to inclusive curriculum. *Center for Educational Policy Studies Journal*. https://doi.org/10.26529/cepsj.543.

Nunally, J. (1978). Psychometric theory *(2^{nd} ed).* New York: McGraw Hill.

Nurhalim, M. (2018). Mengembangkan Format Penilaian Komprehensif dalam Pengembangan KTSP. *INSANIA : Jurnal Pemikiran Alternatif Kependidikan*, 15(3), 414–429. https://doi.org/10.24090/insania.v15i3.1555.

Popham, W. J. (1995). Classroom assessment what teacher need to know. Boston, USA.

Purwanto, (2009). *Evaluasi Pendidikan*, Jakarta: Rineka Cipta.

Retnawati, Heri, (2016). *Validitas reliabilitas & karakteristik butir.* Yogyakarta: Parama Publishing.

Setiawan, J., Ajat Sudrajat, A., Aman, A., & Kumalasari, D. (2021). Development of higher order thinking skill assessment instruments in learning Indonesian history. International Journal of Evaluation and Research in Education (IJERE), 10(2), 545. https://doi.org/10.11591/ijere.v10i2.20796.

Siti Kholidatur Rodiyah. (2019). Ideal Evaluation in Islamic Education Learning. *EDUTEC : Journal of Education And Technology*, 2 (2), 1–5. https://doi.org/10.29062/edu.v2i2.23.

Slepkov, A. D., Van Bussel, M. L., Fitze, K. M., & Burr, W. S. (2021). A Baseline for Multiple-Choice Testing in the University Classroom. *SAGE Open*. https://doi.org/10.1177/21582440211016838.

Sutarno, S., & Hernawati, S. P. (2022). Comparative Study of the Use of Objective Tests and T essays D in Improving Achievement Learning Indonesian Language. *PINISI Discretion Review*, 5(2), 391. https://doi.org/10.26858/pdr.v5i2.32434.

Wainer, H., & Braun, H. I.. (1990). *Tes validity.* New Jersey: Hillsdale 07642.