Advances in Social Science, Education and Humanities Research, volume 261
International Conference on Islamic Education (ICIE 2018)

# The Assumption Testing and Compatibility Model of Final Examination in Islamic Religious Education by Item Response Theory

Lian G. Otaya
State Islamic Institute (IAIN) Sultan Amai Gorontalo
Jalan Gelatik no.1, Gorontalo 96112, Indonesia
lian.otaya@yahoo.com

***Abstract*— This study aimed to examine unidimension assumption, local independency and parameter invariance and also compatibility of model at final examination in Islamic education subject by item response theory. Research Methodology used in this research is quantitative research with expose-facto approach. Data were collected by document study to the students' response on multiple choices on Islamic education subject at XII grade of Senior High School in Gorontalo academic years 2016/2017 with 320 samples. Assumption test data analysis conducted by factor analysis by helping of SPSS version 24.0 software and compatibility statistic test of one item with model or goodness of fit statistic by using BILOGMG program version 3.0. The result shows that unidimension assumption testing gained information about fulfilled measurement model, because it is found many curves of eigen value from the output of total variance which explains the value of eigen in each component and explained variants and also the picture of scree-plot. The test of local independency assumption is automatically proven because the result of response data analysis from students is not unidimension. The appropriate model used to analyse the questions of Islamic education subject at final examination is 2 parameter logistic model (2PL). The assumption of 2PL parameter invariance testing shows that there is no estimation result of parameter variance at odd group and even groups. It means that parameter invariance of index discrimination (a) and difficulty index (b) is fulfilled.**

***Keywords*— *Assumption testing; compatibility model; Item response theory***

## I. INTRODUCTION

Measurement in education is an effort to set systemcatically a number on each student for describing the characteristics of students' abilities in particular fields. According to Allen & Yen, measurement is the assigning of numbers to individual in a systematic way as a means of representing properties of individual [1]. But, there are often errors in measurement. As the measurement through test, the sources of measurement errors are on the determination of exam materials, the side which is measured, the side who measures and the environment [2].

The measurement theory which is recently developed consists of two kinds of estimation method, those are: classical test theory and item response theory (modern theory) or which is more popular to be recognized as Item Response Theory (IRT) [3]. IRT is a family of models that share some fundamental ideas [4].

IRT is said to have two basic assumptions which are unidimensionality and local item independence [5], and in order to determine the appropriate IRT model, first it should be assessed whether the data set at hand meets the IRT assumptions or not [6]. This assumption states that the conditional probability of observing a response pattern given a particular latent trait value equals the product of the items' conditional probabilities [7].

Mathematical model of IRT has a meaning that the subject probability to answer the items correctly depends on the abilities of subject and the characteristics of items [8]. Some assumptions need to be met in order to obtain valid results with the IRT models developed for both dichotomous and polytomous items. These assumptions are unidimensionality, local independence and model data fit [9].

Models developed for the dichotomous items vary depending on the difficulty of item (b), item discrimination (a), and pseudo-guess (c) parameters [10].

Besides the assumptions which have been stated in IRT, the other important things that need to be noticed is the choice of the appropriated model. There are three kinds of IRT logistic model, those are: one parameter logistic model (1PL), two parameters logistic model (2PL), and three parameters logistic model (3PL).

1PL model is a an item parameter estimation model which reviews the level of item difficulties with assumes that the difference point is same for all items and the guess is same with 0.Then, 2PL model is a model which emphasizes except on item difficulty items , and also emphasizes on the difference point of question items. While, 3 PL model is a method in IRT where the difficulty level, difference point, and the guessing are controlled altogether [11].

Observing the opinions above, it is indicating the importance of verification and testing of unidimensional assumptions, local independence, parameter invariance and model suitability based on response theory items to provide empirical evidence in investigating the internal structure of the

constants defined by the test developer, and to find out the extent to which theoretical structure is proven on the response of the test participants. If it is proven, the significance of the score generated by the test contains an interpretation that is consistent with its measurement objective. In addition, the selection of appropriate models will reveal the true state of the test data as a result of measurement, including on the final exam of the subjects of Islamic Religious Education.

## II. METHOD

This research is a quantitative research with expost facto approach. Data was collected by using documentation technique consist of the questions of final semester test at Islamic religious education subjet class XII of of Senior High School in Gorontalo Academic Year 2016/2017 and students' answer sheets from 40 question items. This research population consists of 1.321 students. The sampling technique used *proporsional random sampling* and the determination of the number of samples by using *Nomogram Harry King's* table for the error 5% with trust level 95% and its multiplying factor is 1,195 obtained 320 students as the samples. The data analysis technique of *unideimension* assumption data is conducted through the factor analysis with the helps from SPSS program version 24.0 and for the testing of the suitabilities of model based item response theory logistic model (1PL, 2PL dan 3PL) with the help of BILOG-MG version 3.0.

## III. RESULTS AND DISCUSSION

The finding of this research is obtained the unidimension testing result and the suitabilities of the model. The testing of unidimension is conducted to know whether the test used in measuring one kind of *trait* which is to measure the ability of students at grade XII of Senior High School in Gorontalo at academic year 2016/2017 on the subject of Islamic religious education. Then, the parameter invariance assumption testing and the suitabilities of the model are carried out to analyze the suitabilities of question items with the model or the *goodness of fit statistic*.

### *A. Unidimension Testing*

To examine whether there is correlation between the dimensions, it is used *Barlett's test of sphericity* test. If the result is significant, it means that the correlation matrix has significant correlation with some dimensions. The other tests which is used to see the intercorrelation among variables and whether factor analysis can be conducted or not is by using Kaiser Meyer Olkin *Measure Of Sampling Adequacy* (KMO-MSA). The output results of factor analysis through KMO-MSA test dan Bartlett's test are shown in the table below:

TABLE I. KMO AND BARLETT'S TEST

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .873 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 2.843E3 |
| | df | 780 |
| | Sig. | .000 |

Table 1 shows the value of chi-square on Tabel 1 menunjukkan nilai chi-square on Bartlet's test is 2.843 df 780 with the significance value 0.000. This thing shows that the amount of sample 320 which is used in this research has been enough and there is correlation among dimensions. To the value of KMO *Measure of sampling Adequacy* (MSA) is 0.873 from the data which were analysed ≥ 0.05 so the anlaysis factor is feasible to the furthermore process. According to Ghozali, if the value of MSA shows the ammount ≥ 0.05 so the abalysis factor can be done [12].

The next, to get the items which measure same dimensions, it is conducted the extraction process so it results some factors. Many factors formed were shown by the components which have eigenvalue >1, which can be seen at the table below:

TABLE II. TOTAL VARIANCE EXPLAINED

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 7.928 | 19.820 | 19.820 | 7.928 | 19.820 | 19.820 |
| 2 | 1.929 | 4.823 | 24.643 | 1.929 | 4.823 | 24.643 |
| 3 | 1.664 | 4.160 | 28.804 | 1.664 | 4.160 | 28.804 |
| 4 | 1.422 | 3.556 | 32.359 | 1.422 | 3.556 | 32.359 |
| 5 | 1.310 | 3.275 | 35.634 | 1.310 | 3.275 | 35.634 |
| 6 | 1.298 | 3.246 | 38.880 | 1.298 | 3.246 | 38.880 |
| 7 | 1.168 | 2.920 | 41.800 | 1.168 | 2.920 | 41.800 |
| 8 | 1.142 | 2.854 | 44.654 | 1.142 | 2.854 | 44.654 |
| 9 | 1.134 | 2.835 | 47.489 | 1.134 | 2.835 | 47.489 |
| 10 | 1.123 | 2.808 | 50.297 | 1.123 | 2.808 | 50.297 |
| 11 | 1.058 | 2.644 | 52.941 | 1.058 | 2.644 | 52.941 |
| 12 | 1.047 | 2.618 | 55.558 | 1.047 | 2.618 | 55.558 |

*Extraction Method: Principal Component Analysis*

The result of factor analaysis is there are 12 factors with eigenvalue >1, so it can be said that 40 items analysed is grouped into 12 factors. Those 12 factors explain about 55,558% from the total of varians. The result showed that the first factor explained 19,820% from the total of varians. The eigenvalue from first factor was two much bigger than the eigenvalue of secon factor, so it can be said that those factors had formed dominant factor.

Naga stated that the eigenvalue of first factor is bigger than the igenvalue of the second factor, while the eigenvalue of the second factor and the next is almost same so it can be stated that the requirement of unidimension had been fulfilled [13].

Then, if we see from the matrix component so the number of items collected from the first factor has been dominant, those are 25 items from 40 analysed items (70%). At the total variance output which explained the amount of eigenvalue on the each component and varians which are explained and also the picture of *scree-plot* and it was found many eigenvalue curves, this thing showed that the single measurement assumption model had been fulfilled. This thing can be seen at the *scree-plot* picture as follow:

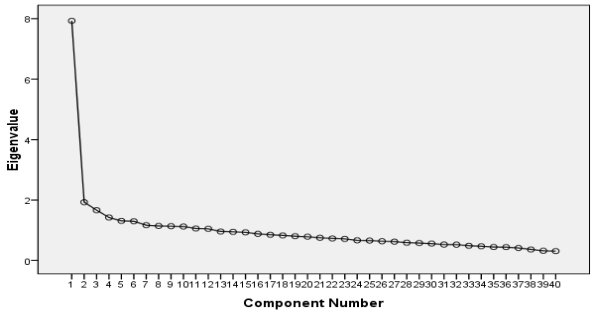


Fig. 1. *Scree-Plot Eigen Value*

Picture 1 shows that there is one curves which is produced by eigenvalue on the first component >1, that is 7,928. The ammount next eigenvalue is down to 1,929 at the eigenvalue of the second component, and the third eigenvalue and the next is relatively small with value <1 which cause the sloping of *scree-plot*. The curve which is formed in the *scree-plot* as many as 1 curve whereas the change of another eigenvalue is formed in slope, it shows that there is only 1 dimension which is measured in that instrument. The second assumption is a local indepence assumption. This assumption is automatically proven.

*B. The Testing of parameter invariance and model suitability*

The final assumption which will be proven that is parameter invariance assumption. Parameter invariance is proven by seeing the item of invariance parameter and the parameter of the test participants' ability. The first step is to break the respondent into two part. In this study, the repsondent is broken into two based on some patterns. The participants who have odd number are grouped and are analysed separately with the participants who have even number. Because the number of respondents are 320, so the number of respondent at odd number and even number for each as many as 160 people of the test participants.

Before the proof from the items of parameter invariance and ability, the first step is to decide the suitability of model which will be used. There are 3 models which can be used, those are 1PL model, 2PL model, and 3 PL model. While, to do an analysis whether the collcted data have bee appropriated with one of the models, there are two ways which can be used, that is with the statistical suitabilities of the model and with the item characteristic of curve plot.

On the statistical choice of model, from the three models are made the item suitabilities based on Chi-quare value. When the logistic model has the most suitable item,so the model will be choosen as the model for the data analysis.

This suitability model testing can be analysed by using BILOG program. After the data anlysis with BILOG program had been conducted, so the output of phase 2 produced item suitability model or *goodness of fit statistic*. The criterion used were if p< *α (item didn't fit with the model) and p ≥ α (item fitted with the model)*.

TABLE III. ITEM SUITABILITY TESTING WITH 1PL,2PL,3PL MODEL

| No | 1PL | | | 2PL | | | 3PL | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Prob/ Sig** | **α** | **Status** | **Prob/ Sig** | **α** | **Status** | **Prob/ Sig** | **α** | **Status** |
| 1 | 0.0250 | 0.05 | Outfit | 0.6857 | 0.05 | Outfit | 0.0677 | 0.05 | Infit |
| 2 | 0.0805 | 0.05 | Infit | 0.4873 | 0.05 | Infit | 0.4744 | 0.05 | Infit |
| 3 | 0.3335 | 0.05 | Infit | 0.6918 | 0.05 | Infit | 0.8752 | 0.05 | Infit |
| 4 | 0.0034 | 0.05 | Outfit | 0.1101 | 0.05 | Infit | 0.3513 | 0.05 | Infit |
| 5 | 0.1018 | 0.05 | Infit | 0.1734 | 0.05 | Infit | 0.5115 | 0.05 | Infit |
| 6 | 0.6368 | 0.05 | Infit | 09802. | 0.05 | Infit | 0.8403 | 0.05 | Infit |
| 7 | 0.0003 | 0.05 | Outfit | 0.0075 | 0.05 | Outfit | 0.0000 | 0.05 | Outfit |
| 8 | 0.0056 | 0.05 | Outfit | 0.0892 | 0.05 | Infit | 0.0388 | 0.05 | Outfit |
| 9 | 0.8037 | 0.05 | Infit | 0.2715 | 0.05 | Infit | 0.0372 | 0.05 | Outfit |
| 10 | 0.5539 | 0.05 | Infit | 0.9845 | 0.05 | Infit | 0.1916 | 0.05 | Infit |
| 11 | 0.6298 | 0.05 | Infit | 0.8434 | 0.05 | Infit | 0.9094 | 0.05 | Infit |
| 12 | 0.0789 | 0.05 | Infit | 0.7676 | 0.05 | Infit | 0.6574 | 0.05 | Infit |
| 13 | 0.8190 | 0.05 | Infit | 0.7513 | 0.05 | Infit | 0.9632 | 0.05 | Infit |
| 14 | 0.0011 | 0.05 | Outfit | 0.0000 | 0.05 | Outfit | 0.0001 | 0.05 | Outfit |
| 15 | 0.0307 | 0.05 | Outfit | 0.0031 | 0.05 | Outfit | 0.7027 | 0.05 | Infit |
| 16 | 0.0273 | 0.05 | Outfit | 0.3405 | 0.05 | Infit | 0.9231 | 0.05 | Infit |
| 17 | 0.6248 | 0.05 | Infit | 0.3077 | 0.05 | Infit | 0.5361 | 0.05 | Infit |
| 18 | 0.0300 | 0.05 | Outfit | 0.0437 | 0.05 | Outfit | 0.0274 | 0.05 | Outfit |
| 19 | 0.0010 | 0.05 | Outfit | 0.2723 | 0.05 | Infit | 0.0957 | 0.05 | Infit |
| 20 | 0.0459 | 0.05 | Outfit | 0.7448 | 0.05 | Infit | 0.4360 | 0.05 | Infit |
| 21 | 0.2030 | 0.05 | Infit | 0.3387 | 0.05 | Infit | 0.0308 | 0.05 | Outfit |
| 22 | 0.5035 | 0.05 | Infit | 0.6008 | 0.05 | Infit | 0.4671 | 0.05 | Infit |
| 23 | 0.0025 | 0.05 | Outfit | 0.1382 | 0.05 | Infit | 0.0742 | 0.05 | Infit |
| 24 | 0.6760 | 0.05 | Infit | 0.8639 | 0.05 | Infit | 0.8554 | 0.05 | Infit |
| 25 | 0.0013 | 0.05 | Outfit | 0.2939 | 0.05 | Infit | 0.3353 | 0.05 | Infit |
| 26 | 0.1002 | 0.05 | Infit | 0.5789 | 0.05 | Infit | 0.8543 | 0.05 | Infit |
| 27 | 0.0000 | 0.05 | Outfit | 0.1327 | 0.05 | Infit | 0.0000 | 0.05 | Outfit |
| 28 | 0.0076 | 0.05 | Outfit | 0.6346 | 0.05 | Infit | 0.9621 | 0.05 | Infit |
| 29 | 0.4345 | 0.05 | Infit | 0.9327 | 0.05 | Infit | 0.3980 | 0.05 | Infit |
| 30 | 0.2396 | 0.05 | Infit | 0.3733 | 0.05 | Infit | 0.6700 | 0.05 | Infit |
| 31 | 0.0041 | 0.05 | Outfit | 0.4706 | 0.05 | Infit | 0.5793 | 0.05 | Infit |
| 32 | 0.6589 | 0.05 | Infit | 0.1005 | 0.05 | Infit | 0.2733 | 0.05 | Infit |
| 33 | 0.0025 | 0.05 | Outfit | 0.3282 | 0.05 | Infit | 0.0169 | 0.05 | Outfit |
| 34 | 0.5365 | 0.05 | Infit | 0.8003 | 0.05 | Infit | 0.2837 | 0.05 | Outfit |
| 35 | 0.1721 | 0.05 | Infit | 0.9262 | 0.05 | Infit | 0.7954 | 0.05 | Infit |
| 36 | 0.0000 | 0.05 | Outfit | 0.0113 | 0.05 | Outfit | 0.0190 | 0.05 | Infit |
| 37 | 0.3082 | 0.05 | Infit | 0.7178 | 0.05 | Infit | 0.4837 | 0.05 | Infit |
| 38 | 0.4965 | 0.05 | Infit | 0.3493 | 0.05 | Infit | 0.4489 | 0.05 | Infit |
| 39 | 0.0198 | 0.05 | Outfit | 0.0162 | 0.05 | Outfit | 0.2144 | 0.05 | Infit |
| 40 | 0.1654 | 0.05 | Infit | 0.0875 | 0.05 | Infit | 0.1997 | 0.05 | Infit |
| Total Infit Model | | | 22 | Infit Model | | 34 | Infit Model | | 31 |
| Total Outfit Model | | | 18 | Outfit Model | | 6 | Outfit Model | | 9 |

Table 3 shows that the model which produces suitable item with the model which is more than 2PL model. This means that 2PL model is a model which can be chosen to the analysis of item characteristics. For the brief understanding about the suitable items with the model, it is made the distribution of item suitability with the model as seen on the table 4 below:

TABLE IV. THE DISTRIBUTION OF ITEM SUITABILITY WITH THE MODEL

| MODEL 1PL | MODEL 2PL | MODEL 3PL |
|---|---|---|
| 2,3, 5,6, 9,10,11,12,13 17, 21,22,23 24, 26 29,30,32,34,35,37 38,40 | 1,2,3,4,5,6,8,9,10,11,12 13, 16,17,19,20,21,22,23 24,25,26,27,28,29,30,31 32,33,34,35,37,38,39,40 | 1,2,3,4,5,6,11,12,13 15,16,17,19,21,22,23 24,25,26,28,29,30,31 32,34,35,37,38,39,40 |
| **Total 22 Item** | **Total 34 Item** | **Total 31 Item** |

From the table 4 above, we can see that 2PL model is the best model if it is compared with 1PL model and 3 PL model. Hence, the most suitable logistic model to measure the item suitability with the model is 2PL model which its results are served in the table below:

TABLE V. ITEM SUITABILITY TESTING WITH 1PL, 2PL, 3PL MODEL

| ITEM | SLOPE | THRESH | ASYMP | CHISQ | PROB | STATUS |
|---|---|---|---|---|---|---|
| 1 | 0.873 | -0.427 | 0.000 | 3.9 | 0.6857 | INFIT |
| 2 | 0.973 | -0.170 | 0.000 | 5.5 | 0.4873 | INFIT |
| 3 | 0.656 | -1.138 | 0.000 | 3.9 | 0.6918 | INFIT |
| 4 | 1.246 | -0.153 | 0.000 | 9.0 | 0.1101 | INFIT |
| 5 | 1.201 | -0.139 | 0.000 | 7.7 | 0.1734 | INFIT |
| 6 | 0.811 | -0.134 | 0.000 | 1.6 | 0.9802 | INFIT |
| 7 | 1.214 | -0.848 | 0.000 | 0.13.9 | 0.0075 | OUTFIT |
| 8 | 0.954 | 0.066 | 0.000 | 0.11.0 | 0.0892 | INFIT |
| 9 | 0.626 | 0.066 | 0.000 | 0.8.7 | 0.2715 | INFIT |
| 10 | 0.931 | -0.477 | 0.000 | 1.0 | 0.9845 | INFIT |
| 11 | 0.777 | -0.107 | 0.000 | 2.7 | 0.8434 | INFIT |
| 12 | 0.829 | 0.347 | 0.000 | 4.1 | 0.7676 | INFIT |
| 13 | 0.523 | -0.526 | 0.000 | 4.2 | 0.7513 | INFIT |
| 14 | 0.440 | 1.187 | 0.000 | 36.4 | 0.0000 | OUTFIT |
| 15 | 0.564 | 0.178 | 0.000 | 21.5 | 0.0031 | OUTFIT |
| 16 | 0.891 | 0.212 | 0.000 | 7.9 | 0.3405 | INFIT |
| 17 | 0.715 | 0.151 | 0.000 | 9.4 | 0.3077 | INFIT |
| 18 | 0.429 | 0.247 | 0.000 | 15.9 | 0.0437 | OUTFIT |
| 19 | 1.231 | -0.428 | 0.000 | 6.4 | 0.2723 | INFIT |
| 20 | 0.413 | 0.378 | 0.000 | 5.1 | 0.7448 | INFIT |
| 21 | 0.948 | 0.137 | 0.000 | 7.9 | 0.3387 | INFIT |
| 22 | 0.449 | 0.347 | 0.000 | 6.4 | 0.6008 | INFIT |
| 23 | 0.501 | 0.362 | 0.000 | 11.0 | 0.1382 | INFIT |
| 24 | 0.998 | -0.129 | 0.000 | 2.5 | 0.8639 | INFIT |
| 25 | 0.351 | 1.263 | 0.000 | 9.6 | 0.2939 | INFIT |
| 26 | 0.837 | -0.349 | 0.000 | 4.7 | 0.5789 | INFIT |
| 27 | 0.195 | 1.425 | 0.000 | 12.4 | 0.1327 | INFIT |
| 28 | 0.299 | 2.799 | 0.000 | 6.1 | 0.6346 | INFIT |
| 29 | 0.715 | -0.902 | 0.000 | 1.3 | 0.9327 | INFIT |
| 30 | 0.660 | -0.164 | 0.000 | 6.5 | 0.3733 | INFIT |
| 31 | 0.293 | 0.503 | 0.000 | 7.6 | 0.4706 | INFIT |
| 32 | 0.539 | 0.140 | 0.000 | 13.3 | 0.1005 | INFIT |
| 33 | 0.266 | 1.465 | 0.000 | 9.2 | 0.3282 | INFIT |
| 34 | 0.860 | -0.313 | 0.000 | 3.1 | 0.8003 | INFIT |
| 35 | 0.855 | 0.063 | 0.000 | 1.9 | 0.9262 | INFIT |
| 36 | 0.245 | 2.194 | 0.000 | 18.2 | 0.0113 | OUTFIT |
| 37 | 0.520 | -0.662 | 0.000 | 4.5 | 0.7178 | INFIT |
| 38 | 0.526 | 0.233 | 0.000 | 8.9 | 0.3493 | INFIT |
| 39 | 0.349 | 0.149 | 0.000 | 18.8 | 0.0162 | OUTFIT |
| 40 | 0.930 | -0.645 | 0.000 | 9.6 | 0.0875 | INFIT |

Table 5 above shows the number of question items which is information estimacy *measure of difficulty*, *standard error of measurement*, the data suitability with the model (infit and outfit), and also the correlation of question item's difference point (*point bisserial*) from 2PL model which is more than 1PL and 3PL model.

The result analysis of 2PL model, then it is used to give information about parameter invariant based on the the grouping of odd number students ana even number students. The data of each group then it will be analysed its item parameter. In this case, because of the most suitable model to be used is 2PL model, so the estimation item parameter which is stated in the form of difference point (a) and difficulty level (b).

The parameter aand b from each group is estimated. For the parameter of odd number student group and even number student group are described in a scatter diagram *(seater plot)*. So do with parameter b.

Scatter diagram for the difference point, which is estimated on the odd number student and even number student, which are described on picture 2

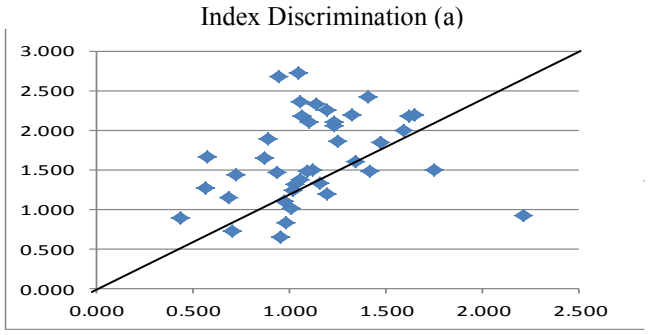


Fig. 2. Invariance Parameter at odd number student and even number student

Based on that picture, it is obtained that each point is relatively near on the line with the slope 1 (the equivalence of line y y = x). This thing shows that there is no parameter variance of estimation result at the odd number student group and even number student group. In other words, parameter invariance (a) is fulfilled. Scatter diagram for the difficulty level, which is estimated on odd number student group and even number student groupare described at picture 3.

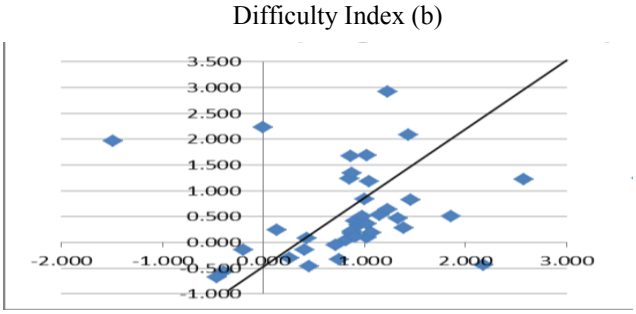


Fig. 3. Parameter Invariance odd number student group and even number student group

Based on the picture of *scater-plot* with the parameter of difficulty level which has been compared with the straight line which has gradient 1(the equivalence of line y= x), it looks that the difficulty level of point ordinate from each item which is realtively near on the straight line. This thing shows that there is no parameter invariance of estimation result at odd number student group and even number student group. In the other words, parameter invariance (b) is fulfilled.

The next is the result of item parameter estimation with 2PL model stated in the form of difference point index (a) and difficulty level (b), it is explained as follow:

*1) Index Discrimination*

Parameter $a_i$, is an index discrimination which is had by the first item. There is characteristic curve, $a_i$ which is proportional to the coefficient of the direction of intersect line (*slope*) at the point θ= b. Question item with big index discrimination has very ascending curve, whereas the question items with a small index discrimination which have very slope curve. Theoritially, the value of $a_i$ is located between $-\infty$ and $+\infty$. On the good item, this value has a positive relationship with performance on the item with the measured item, and $a_i$ is located between 0 and 2 [11].

Index discrimination from output BILOG-MG can be seen from output phase 2. The classification of item difference point index is served as follow:

TABLE VI. THE RESULT OF PARAMETER DISCRIMINATION INDEX ANALYSIS

| NO | Index discrimination ($a_i$) | Category | Number of Items | Item Number |
|---|---|---|---|---|
| 1 | $a_i > 2.0$ | Less | 7 | 25,27,28,31,33,36,39 |
| 2 | $-2.0 \leq a_i \leq 2.0$ | Good | 33 | 1,2,3,4,5,6,7,8,9,10,11 12,13,14,15,16,17,18 19,20,21,22,23,24,26 29,30,32,34,35,37,38,40 |
| Total Item | | | 40 | |

Table 6 shows that about 82,5% question item has a good item difference point. This thing proves that 82,5% from all items of final test's questions at the subject of Islamic religious education at grade XII of Senior High School in Gorontalo academic year 2016/2017 has ability to assert the differences among the participants of the test who can answer the questions correctly and who answer the questions incorrectly.

*2) Difficulty Index*

Parameter $b_i$ is a point at the ability skill so the chance to answer correctly is 50%. The bigger the value of $b_i$, the bigger the ability which is needed to answer the questions correctly with chance as 50 % or in the other words the bigger the value of parameter $b_i$, so the harder those question items. The value of $b_i$ (difficulty index) is various from -2 untill +2. The value near to -2 shows that those items are very easy whereas the value near to +2 shows that those items are very difficult [11].

Difficulty index from ouput BILOG-MG can be seen on Threshold value at the output of phase 2. Based on the result of analysis data, it was obtained the information of difficulty level of all items which moved from the value -1.138 until 2.799. The classification of difficulty index item is presented as follow:

TABLE VII. THE RESULT OF PARAMETER DIFFICULTY INDEX ANALYSIS

| NO | Difficulty index ($b_i$) | Kategori | Jumlah Item | Nomor Item |
|---|---|---|---|---|
| 1 | $b_i > 2.0$ | Kurang Baik | 2 | 28, 36 |
| 2 | $-2.0 \leq b_i \leq 2.0$ | Baik | 38 | 1,2,3,4,5,6,7,8,9,10,11 12,13,14,15,16,17,18 19,20,21,22,23,24,26,27 29,30,31,32,33,34,35,373 8,39,40 |
| Total Item | | | 40 | |

Based on the table 7, it was gained the information that about 95% from all items have good difficulty level. This information proves that about 95% from the all test items is able to describe the function of someone's ability, where the testee who has good ability will feel easy to do question items. In contrary, the testee who has bad ability will feel difficult to answer the question items. While 5% of the rest is included to the difficult item because it has the difficulty level which is more than 2.0. If the test is quite difficult, it means that a half of students' ability level tends to be weak.

Based on the analysis of fit model and the estimation of the parameter above, so it can be decided how many good items which fulfill the IRT criteria. To know empirically the quality of the item, it can be used the regulations from Hulin, Drasgow and Parsons with criteria good, less, and bad [14]. The criteria of each parameter as follow:

a) Item is in good category when the item is suitable with the model, has difficulty index from -2.0 until 2.0, difference point index of item 0.0 until 2.0, and the chance of guessing is not more than 0.2

b) Item is in less category when the item is suitable with the model, has difficulty index which less than -2.0 or more than 2.0,difference point index of item is more than 2.0,and the chance of guessing is more than 0.2.

c) Item is I bad category if it is not suitable with the model.

Based on 40 items of final semester tetst at subject of Islamic Religious Education at Grade XII of Senior High School at Gorontalo in academic year 2016/2017, so there are 6 items or about 15% item which has category less, they are: question item number 7, 14, 15, 18, 36, and 39. Those six question items didn't fulfill the qualification of IRT analysis, so those six items are not good items (bad).

The all question items of final semester test at the subject of Islamic Religious Education at Grade XII of Senior High School at Gorontalo in academic year 2016/2017, there is 85% item which has good characteristic/good quality, where the items fit the model and fulfill the estimation of difference point parameter, the good difficulty level and pseudo guessing. While, 15% of question numbers is in category good

because there is one or more criteria of item parameter estimation which doesn't fulfill the criteria of god item, whether that is dicrimination index parameter and difficulty index.

## IV. CONCLUSION

The testing of unidimension assumption based on the item of response theory shows that the outpu of ttotal *variance* which explain the ammount of eigenvalue at each component and variants which are explained and also the picture of *scree-plot* which is found many eigenvalue curves, this thing show that the assumption of unidimension measurement model is fulfilled. The second assumption is the local independence assumption. This assumption is automatically proven because the result of data analysis response of students is multidimension and not unidimension.

The most suitable model to be used to analyze the result of final semester test at the subject of Islamic Religious Education Grade XII of Senior High School at Gorontalo is the 2PL logistic model.so the assumption of 2 PL invariance parameter testing is obtained the information from the scatter diagram *(seater plot)* where each point is relatively near to the line with the slope 1 (the equivalence of line y = x). This thing shows that there is no parameter variation of estimation result at the odd number student group and the even number student group. In the other words, the parameter invariance of difference point index (a) and the difficulty level (b) is fulfilled.

## REFERENCES

[1] Allen, M.J & Yen, W.M, *Introduction Measurement Theory*. Brooks/Cole. Publishing Company, 1979.

[2] Mardapi, D, *Pengukuran, penilaian dan evaluasi pendidikan*. Yogyakarta: Parama Publishing, 2016.

[3] McDonald, R.P, *Test theory: A unified treatment*. New Jersey: Larvrence Erbaum Associates, Publishers, 1999.

[4] De Gruijter, D. N., & van der Kamp, L. J. T, *Statistical test theory for the behavioral sciences*. CRC Press, 2007.

[5] Hambleton, R. K., & Swaminathan, H, *Item response theory: Principles and applications*. USA: Kluwer-Nijhoff Publishing, 1985.

[6] Koğar, E.Y., & Kelecioğlu, H, "Examination of different item response theory models on tests composed of testlets," Journal of Education and Learning, Vol. 6 No. 4, 2017, p. 113-125.

[7] Liu, Y., & Olivares, A.M, "Local dependence diagnostics in IRT modeling of Binary Data," Educational and Psychological Measurement 73(2), 2012, p. 254-274.

[8] Retnawati, H, *Teori respon butir dan penerapannya*. Yogyakarta: Parama Publishing, 2014.

[9] Aybek, E.C. & Demirtasli, R.N, "Computerized adaptive test (CAT) applications and item response theory models for polytomous items," International Journal of Research in Education and Science (IJRES), 3(2), 2017, p. 475-487. DOI: 10.21890/ijres.327907.

[10] DeMars, C, *Item response theory*. Oxford: Oxford University Press, 2010.

[11] Hambleton, R.K., Swaminathan H. & Rogers, H.J, *Fundamental of item response theory*. Newbury Park, CA: Sage Publication Inc, 1991.

[12] Ghozali, I, *Aplikasi analisis multivariat*. Semarang: Universitas Diponegoro, 2006.

[13] Naga, D.S, "Karakteristik butir pada alat ukur model dikotomi," Arkhe: Jurnal Ilmiah Psikologi, III (4), 1998, p. 34-42.

[14] Hullin, C.L., et al, *Item response theory: Application to psichologycal Measurement*. Homewood, IL: Dow Jones-Irwin, 1983.