# The Exploration Of Non-Linear Model was Reviewed Based On The Length Of The Test

**Ode Zulaeha[1*], Wardani Rahayu[2] , Yuliatri Sastrawijaya[3]**

[1] STKIP Kie Raha Ternate, Indonesia

[2, 3]State University of Jakarta, Indonesia

* zulaehapepunj@gmail.com

**ABSTRACT**

The purpose of this study is to explore the resulting dichotomous nonlinear data model analysis on a multidimensional three-parameter logistic model. This study used a sample size of 1500 with grain lengths of 20 items, 30 items, and 40 items. The instrument used was the 2015 DKI Jakarta National Examination Question. The data used were analyzed using testfact software. The results showed that the use of the nonlinear model in the multidimensional three-parameter logistic model was very good, because the participants' low ability in other dimensions would answer correctly in other dimensions.

**Keywords**

non linear model, multidimensional item response items, and tetrachoric correlation

*Article Received: 18 October 2020, Revised: 3 November 2020, Accepted: 24 December 2020*

## Introduction

Efforts to improve the quality of education can be seen from improving the quality of learning and the quality of assessment. Assessment is a process carried out in order to monitor the learning progress of students as evaluation material for future learning improvements. The results of the assessment are presented in numbers and letters as a sign to determine the extent to which learners have mastered a subject matter. Assessment is one of the most important elements in learning and is a component that is no less important than the learning model or method. The quality of the assessment is closely related to the test items used to measure students' abilities, both in the form of knowledge and skills. The test is a systematic procedure for observing or describing one or more characteristics of a person using a numerical standard or categorical system (Cronbach, 1984). Tests can also be used to measure the amount of knowledge that individuals acquire from a limited level of subject matter (Bruce, 1978). The test is one of the most commonly used measurement tools in education and psychology. In practice, tests should be based on objective, transparent, accountable and non-discriminatory. A test kit should only be unidimensional, which means that each test item measures only one ability. Assumptions can only be demonstrated if the test contains one factor that measures the performance of a subject.

In order to obtain high quality instruments, apart from theoretical analysis, it is also necessary to conduct empirical item analysis. Broadly speaking, this empirical item analysis can be divided into two, namely the classical test theory approach and the item response theory (Item Response Theory, IRT). Allen and Yen (1979) explain that the classical test theory or the so-called pure classical score theory is based on an additive model, namely the observation score and the error score. Where is what is meant by the observation score is the sum of the actual score whereas, measurement error is an unsystematic or random error. There are several assumptions in classical test theory. The first assumption is that the measurement error score does not interact with the actual score. The second is that the error score does not correlate with the actual score and the error score on other tests for the same test taker. Third, is the average of these error scores equal to zero.

This classical test theory has several weaknesses. The first, the level of difficulty and distinguishing power of the questions really depends on the sample used in the analysis. The average level of ability, range, and distribution of students' abilities as the sample in the analysis greatly influenced the statistical value obtained. Second, the scores obtained by students from a test are very limited to the tests used so that they cannot be generalized outside of the tests used. Third, the consistency / reliability of the test on the alignment of the test sets is very difficult to fulfill. Fourth, it does not provide a basis for determining how test takers will respond when given certain items. Fifth, the standard error index of measurement is assumed to be the same for each test taker and the sixth is the test item bias test and test equivalence is impractical and difficult to do. To overcome this problem, the item response theory approach was used (Item Response Theory, IRT).

One of the assumptions that must be met in the IRT is unidimensional. Unidimensional assumptions, in some cases the whole item to measure the same ability (Yanyan Sheng and Wikle, 2007). However, Folk and Green (1989) argue that what happens is that many tests measure more than one ability (multidimensional). From the violation of the unidimensional assumption, the test items with more than one dimension can lead to bias in the parameter estimation and result in incorrect parameter estimates. Thus, it is necessary to develop an IRT, namely Multidimensional Item Response Theory (MIRT) Bock and Aitkin, (1981); Reckase and McKinley, (1982); Samejima, (1974); Thissen and Steinberg, (1984); Whitely, (1980).

There are two models in MIRT, namely compensatory and non-compensatory models. According to Spray, Davey, and Reckase, (1990), the compensatory model is a high ability on one dimension which is obtained by compensation for low ability in another dimension in relation to the probability of answering correctly. Meanwhile, the noncompensatory model does not allow high ability in one dimension to get compensation for low ability in other dimensions. For the two-dimensional case compensatory model, a test taker with very low ability in one dimension and very high ability in another dimension can answer the test items correctly. So, this study will apply a three-item compensatory model (Reckase, et al., 1991).

MIRT can be applied to measure general abilities or certain psychological abilities of test takers if the test is multidimensional (Segall, 2000). MIRT can be used for item selection, in order to predict learning and estimate the ability of students. Testing the usefulness of the proposed multidimensional model was carried out to analyze Reckase's (1997) goodness of fit. There is another term for multidimensional IRT, namely the non-linear model. The non-linear model is a test model that differentiates people's abilities, where the easy test for high-ability people and a difficult test for low-ability people (Yalcin, 1995). To prove the assumption of the item response theory on linear and nonlinear multidimensional factors analysis is necessary. There is a very close relationship between the factor analysis model (AF), IRT and MIRT. McDonald (1982) categorized IRT as a non-linear factor analysis group. There are two types of nonlinear factor analysis coupled to IRT and MIRT. First, it is based on the bivariate information method (Bartholomew, 1985; McDonald and Mok, 1995). Second, it is based on comprehensive information methods (Bock and Aitkin, 1981; Bock et al., 1988). Several studies on the multidimensional model of IRT for dichotomous variables such as, (Bock & Aitkin, 1981; Bock & Lieberman, 1970; Lieberman, 1970; McDonald, 1985; Mulaik, 1972; Rasch, 1961; Reckase, 1973; Sympson, 1978; Whiteley, 1980). In general, dichotomous models can be classified into compensatory models, which allow high proficiency in one dimension to compensate for low proficiency in another dimension.

There have been many studies on nonlinear or MIRT models. However, most of them use a small sample size and a multidimensional one-parameter logistic model (Rasch model) or two logistical parameters. Based on previous studies on nonlinear model analysis, it shows that item estimation and the ability of test takers need to use a multidimensional three-parameter logistic model (M3PL) with a larger amount of data as suggested by Rose Marie Batley (1989), Fraser and McDonald (1988). ). The nonlinear model is a model that measures different abilities in answering tests. This model can also measure the stability of item parameters which are influenced by the number of participants (Segall, 2000; Heri, 2008) and the length of the test (Cohen, Kane and Kim, 2001; Heri Retnawati, 2008). So that in this study the M3PL model will be used with a sample size of 1500 and analyzed using testfact software.

## Literature Review

The concept of modern response theory is that the performance of a subject on an item can be predicted or explained by a set of factors called traits, latent traits or ability (ability) and the relationship between the subject's performance on an item and a set of latent abilities that underlie it can be described by a function that increases gradually. monotony is referred to as the item characteristic curve (Hambleton Swaminathan and Rogers, 1991).

Modern item analysis is a review of the items using the IRT or item answer theory. IRT is one way to assess item feasibility by comparing the average number of items against the group's ability predicted by the model. Hambleton & Swaminathan (1985) and Hambleton, Swaminathan, & Rogers (1991) state that there are three assumptions underlying the item response theory, namely unidimensionality, local independence, and group invariance. Unidimensional, meaning that each test item measures only one ability. For example, in the mathematics achievement test, the items constructed are in the form of story questions and dichotomous. If the test taker gives an incorrect response, it cannot be known whether the error was caused by the test taker's imbalance in mathematics or language. In reality it is difficult to come up with an item that measures only one ability of the test taker.

A set of items on a test can be called unidimensional if the test taker's performance can be explained by a latent attribute (Hambleton & Rovinelli, 1986). As a result, it is no longer known the contribution of each ability to the test taker's answers. By changing test items or groups of test takers, invariance can no longer be maintained on test item size and on test taker characteristic measures, so that the inability to maintain this invariance requirement would be contrary to the objectives of the IRT. If a unidimensional requirement has been fulfilled, then a way to determine whether a test item is unidimensional or not is needed, then the factor analysis method can be used. In this case, factor analysis aims to show which factor groups are different items. Each factor can only represent one dimension of the test indicator. So, the most important thing in unidimensional assumptions is the existence of one dominant component that affects the subject's performance.

The probability of answering correctly on a particular item is only affected by the item parameter and a latent attribute ($\theta$). This is what is called the principle of local independence (local independence) (Lord, 1980). The assumption of local independence is that the ability that affects test performance is made constant, so the subject's response to any item will be statistically independent. The assumption of local independence is divided into two, namely local independence to the test taker's response and local independence to the test items (James J. Allen and Yen, 1989). Local independence to the test taker's response means that the correctness of the test taker in answering an item is not influenced by whether or not the other test taker answers the item. Whereas local independence to items means that the response of the subject to one item has no effect on the responses to other items or in simple terms it can be said that the assumption of local independence will be fulfilled if the participant's answer to a question does not depend on the participant's answer to another item. If a latent attribute is not sufficient to explain, then local independence cannot be fulfilled (Stout, 1984, 1989, 2002). Thus, the assumption of local independence refers if the abilities that affect test performance are made constant, then the subject's response to any item will be statistically independent.

The third IRT assumption is invariance. The assumption of invariance is that the grain characteristic curve must correctly reflect the relationship between the unobservable and the observed variables. In IRT, the chances of a student's correct answer, item characteristics or parameters, and test taker characteristics or parameters are linked through a model formula that must be adhered to by both the test item group and the test taker group. This means that the same item for different test takers must follow a predetermined formula or the same test taker for different test items must also follow the formula. This process is called invariance between test items and test takers. In modern measurement, item hardness level is not directly related to respondent's ability. The fundamental difference between classical and modern measurements lies in the invariance of the scoring. In modern scoring, it is invariance (unchanged or fixed) to the test items and to the test takers. According to Lord (1990) that the invariance of the test item parameters through the test-taker group is the most important characteristic of IRT. So, invariance is the test score does not change with respect to the test items as well as for the test taker.

Efforts are being made to develop test kits objectively, from improving test item writing techniques, test administration, to assessing the impact of test length and sample size. Research conducted by Hambleton and Cook (2009) using simulation data revealed the effect of sample size and test length on the stability of item parameter estimates and the ability of test takers on IRT for both 1 parameter, 2 parameters and 3 parameters. The results of the study by measuring the sample size and test length greatly affect the stability of the estimate. Thus, to measure the stability of the estimation in this study, namely testing the effect of, among others, sample size and test length. If the response data which is multidimensional like that is then treated as unidimensional data, it means that it has deviated from the unidimensionality assumption in the UIRT and is also not in accordance with the structural aspects of the measured construct (Messick, 1995). Ackerman (1994) asserts that if an item is imposed unidimensional, the resulting score will be invalid. With the unidimensional requirement, we need a way to determine whether the item is unidimensional or not. One way that is obtained is by using factor analysis. When there is a violation of the unidimensional assumption, then the test items with more than one dimension can lead to bias in the parameter estimate and result in an incorrect parameter estimate. According to Ackerman (1989), Cheng, Wang, and Ho (2009), DeMars (2006), Dirir and Sinclair (1996), Oshima and Miller (1990, Reise, Moore and Haviland (2010, and Yao (2011)) stated that IRT at first it was based on unidimensional assumptions that experienced problems scoring on multidimensional tests. Folk and Green (1989) emphasized that the reality that occurs in the field is that many tests measure more than one ability. Of the several violation factors that occur in item which is unidimensional, one can realize how important dimensions are in a test. According to Abedi (1997) and Kahraman and Thompson (2011) item scores, data analysis and results reports are very influential. So, tests that are MIRT are very important to study. Samejima , (1974); Whitely (1980); Bock and Aitkin, (1981); Reckase and McKinley, (1982); Thissen and Steinberg, (1984); Reckase (1985); Reckase & Ackerman, (1986) stated that it is necessary to develop a multidimensional model IRT l Item Response Theory (MIRT). Spray, et al., (1990); Bolt and Lall (2003) describe the linear logistics MIRT model as:

$$P\,(X_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j, \gamma_j) = \gamma_j + (1-\gamma_j)\,\frac{\exp\,(\propto_j \theta_i + \beta_j)}{1 + \exp\,(\propto_j \theta_i + \beta_j)}$$

Where, $P\,(X_{ij} = 1 \mid \theta_i, \alpha_j, \beta j, \gamma_j)$ is the probability of the examinee answering item j correctly; $X_{ij}$ is the response of the examinee i to item j (0 = false, 1 = true); θi is a vector of the ability of the examinees; $\propto_j$ is the grain parameter vector associated with the grain discrimination strength; $\beta_j$ is a parameter related to grain difficulty; γj is the pseudo-guessing parameter of the item; *i* = 1 ,. . . , I (total number of examinees); and j = 1 ,. . . , J (total number of grains). (De La Torre and Patz, 2005). The parameters of this model include the parameters of the test taker, the power of difference, the level of difficulty and false guesses. The test taker parameters in this model are represented by the elements of the vector $\theta_j$. The number of elements of this vector is something that is often debated (Reckase, 1997). Based on the experience of Reckase and Hirsch (Reckase, 1997), many dimensions of ability are often underestimated or overestimated and this will be detrimental. The number of dimensions used in the model depends on the interaction of items with test takers which need to be adjusted according to the objectives of the analysis.

Another advantage of the multidimensional item response theory according to De La Tore and Patz (2005) is that analysis with the multidimensional item response theory provides additional information that increases the accuracy of item parameter estimation. In multidimensional conditions forming unidimensional conditions, the correlation coefficient between latent variables will be equal to zero. Li, Schafer, & DeMars (2005) reinforce this statement by showing that the multidimensional item response theory approach produces a more accurate score than using the unidimensional item response theory approach.

The linear model estimation in question is the low level of one or more capabilities that can be compensated for on other dimensions. Because compensation is a characteristic of linear combinations, this model is named the linear MIRT model (Spray, et al., 1990; Bolt and Lall, 2003). Meanwhile, the nonlinear model is a form of relationship between response variables and explanatory variables that are not linear in parameters. In general, what is meant by linear and nonlinear models is a test item that is multidimensional. Multidimensional tests began to be used in tests that measure various abilities with the name multidimensional item response theory (MIRT) since the end.

To estimate item parameters and ability parameters, use a multidimensional 3-parameter logistic model (M3PL) which consists of parameter a (discriminant / slope) or the level of steepness of the item value and the ability to answer the item, parameter b (difficulty / intercept / threshold) or the level of difficulty. item, and parameter c (pseudogessing) or pseudo guess. The number of samples used was 1500.

## Methods

This research includes quantitative research. The population in this study were all junior high school students in DKI Jakarta. The question code used was POC5501 2015. The research sample was 1500 respondents with 40 items of tests used. This study uses participant response data at the 2015 *State Junior High School National Examination in Mathematics from the Center for Educational Assessment* (PUSPENDIK) for all areas of DKI Jakarta. The use of MIRT's 3-parameter MIRT compensatory logistic model allows high capabilities on one dimension to be compensated for low abilities in other dimensions in relation to the probability of answering correctly (Spray, Davey, Reckase, et al., 1990).
The actual parameters for slope and threshold are generated according to the following uniform distribution.

**Table 1.** Uniform distribution for slope and threshold

| Parameter | Minimum | Maximum |
|---|---|---|
| a | 1.00 | 2.00 |
| b1 | -2.00 | -1.00 |
| b2 | -1.00 | 0.00 |
| b3 | 0.00 | 1.00 |
| b4 | 1.00 | 2.00 |

### Data analysis

The data analysis technique used the testfact software application. There are three main estimates of threshold, loading, and correlation, which will be observed. The threshold and loading will be converted to the appropriate intercept and slope. The estimation procedure in the program is Maximum Log Likelihood. The estimation of item parameters and capabilities in this nonlinear model only uses a three-parameter logistic model with a sample size of 1500 and item lengths which are categorized into three, namely 20 items, 30 items, 40 items. For the accuracy procedure for the estimation to be executed, the square root of the mean value of variance.

## Results

The results of the nonlinear model MIRT factor analysis are described based on the histogram of the analysis using testfact software. The number of items used in this study were 20 items, 30 items, and 40 items. Where each image will be explained the difference in mean and variance. Like the following figure 1.

Figure 1 factor analysis with a sample size of 1500 and a test length of 20 items shows that the level of student ability is higher, most students answer items correctly. This can be seen by the mean value of the tetrachhoric correlation of 0.2443 and the standard deviation of 0.2877. If made in the form of a curve, the size of the slope or skewness of the data distribution value is positive. If the slope value approaches 0 or 0, then the curve tends to be symmetrical.

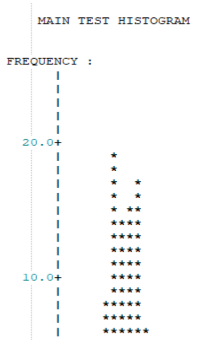

**Figure 1.** Non adaptive data plot full information item analysis factor counted response patterns sample size 20

Spiegel and Stephens (2008). Furthermore, the histogram for a sample size of 1500 and grain length of 30 is as shown in Figure 2 below.

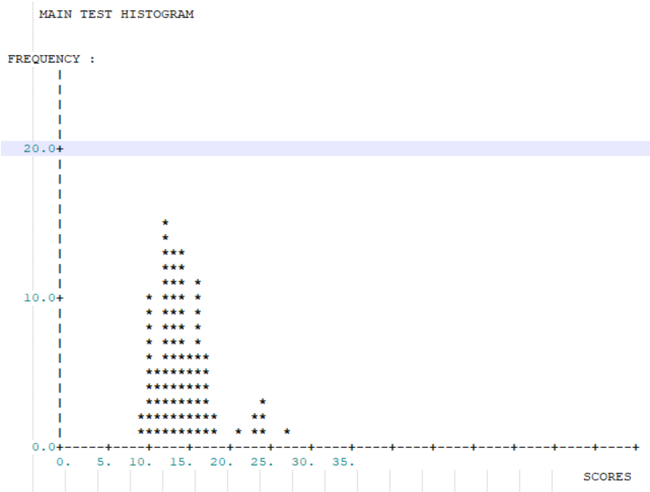


**Figure 2.** Non-adaptive data plot full information item analysis factor counted response patterns sample size 30

Figure 2 factor analysis with a sample size of 1500 and a test length of 30 items shows that the proportion of students who answered correctly was 0.481 with a standard deviation of 0.5. The results of the item factor and ability analysis showed that when participants were given a test of 30 items, the average correlation between the items was 0.1579 and had a standard deviation of 0.25. If it is made in the form of a curve, the size of the slope or skewness of the data distribution value is still skewed to the left, which means it is positive. If the slope value approaches 0 or 0, then the curve tends to be symmetrical. In the use of the test items 30 items the percentage of variance for the slope level is 9.98 with a threshold of 5.017.

From the factor analysis that has been used for each different item, it has a different difference. In order to see the results of this study, the analysis will be described with a sample size of 1500 and a grain length of 40. From the results of the analysis, each difference can be seen. The following describes a sample size of 1500 and a grain length of 40.

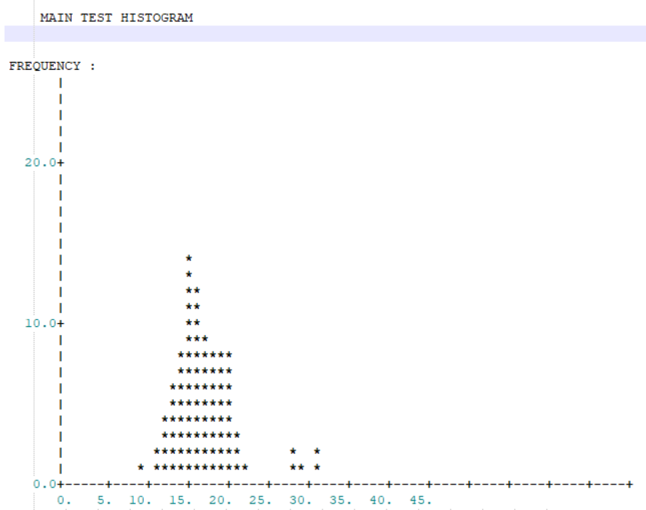


**Figure 3**. Non-adaptive data plot full information item analysis factor counted response patterns sample size 40.

Based on the results of factor analysis using testfact software, it can be seen that there are several items that were discarded. These items are point 1 and point 2. After using the sample size of 40, it can be seen that there is an imbalance which can be seen in table 2 below. Where in items 4, 5, 6, 7, 8, 9, 10, 11, and item 14 have varied root mean square error values. The percentage of variance on the slope is 10.85 and the threshold is 4. 644. This shows that if there are more items, the level of answering with guesses greatly affects the value in the factor analysis.

**Table 2.** Estimation of parameters using testfact software

| No | Cycle | Marginal log likelihood | Change | Intercept | Slope | RMSE |
|---|---|---|---|---|---|---|
| 1 | C3 | 0.4720 | **0.2360** | 0.869 | 0.5035 | **0.2077** |
| 2 | C4 | 0.4718 | 0.1085 | 0.4779 | 0.3246 | 0.1875 |
| 3 | C5 | 0.4716 | 0.6037 | 0.3954 | 0.2866 | 0.1875 |
| 4 | C6 | 0.4716 | 0.3823 | 0.4206 | 0.2982 | 0.1659 |
| 5 | C7 | 0.4715 | 0.2586 | 0.4653 | 0.3186 | 0.1598 |
| 6 | C8 | 0.4715 | 0.1825 | 0.5022 | 0.3351 | 0.1561 |
| 7 | C9 | 0.4715 | 0.1825 | 0.5286 | 0.3469 | 0.1557 |
| 8 | C10 | 0.4714 | 0.1022 | 0.5482 | 0.3558 | 0.1605 |
| 9 | C11 | 0.4714 | 0.8156 | 0.5655 | 0.3641 | 0.1727 |
| 10 | C12 | 0.4714 | 0.6809 | 0.6711 | 0.4281 | **0.2016** |

| No | Cycle | Marginal log likelihood | Change | Intercept | Slope | RMSE |
|---|---|---|---|---|---|---|
| 11 | C13 | 0.4714 | 0.594 | 45.658 | 36.105 | **0.2592** |
| 12 | C14 | 0.4719 | 0.2342 | 25.000 | 10.000 | 0.0615 |
| 13 | C15 | 0.4714 | 0.2327 | 0.1406 | 0.0625 | 0.0635 |
| 14 | C16 | 0.4714 | 0.1498 | 0.2207 | 0.1582 | 0.0632 |
| 15 | C17 | 0.4714 | 0.4217 | 0.3106 | 0.1911 | 0.0623 |
| 16 | C18 | 0.4714 | 0.3424 | 0.0553 | 0.0325 | 0.0611 |
| 17 | C19 | 0.4713 | 0.3103 | 0.7777 | 0.4813 | 0.0596 |
| 18 | C20 | 0.4713 | 0.2861 | 0.5000 | 0.5937 | 0.0581 |

Analysis of variance has an effect on the significance using alpha 0.05 is presented in table 2. From the table above, it can be seen that the sample size of 1500 and the test length of 40 have an average root mean square error which is not good convergence. This is like the case studied by Bock and Zimowski (1999). However, there are three items that have good convergence values, namely items 1, 10, and 11. The level of grain difficulty is higher, then the biserial value is lower.

In general, the tetrachoric grain correlation matrix is not necessarily positive. This means that they often cannot be used in any of the many statistical procedures that require positive certainty, such as calculating partial correlations among some items while keeping others fixed.

Table 3 shows that the length of the test 40 has a greater S.D value, so the data distribution tends to be far from the average. This can be seen in every image. For the length of the test 20 the ability of the test taker is higher or most of the test takers can almost correctly answer the items. Likewise for the test length of 30. It is different from the length of the 40 test. The tetrachoric correlation has an unstable estimate of the item length of 20 because the test items have a frequency close to 0

**Table 3.** Non-adaptive Information Item Factor Analysis on tetrachhoric correlation

| Test lenght | S.D | Corelation | Variansi (%) | |
|---|---|---|---|---|
| | | | 1 | 2 |
| 20 | 0.25 | 0.24 | 10.98 | 3.75 |
| 30 | 0.27 | 0.15 | 9.98 | 5.01 |
| 40 | 0.28 | 0.19 | 10.8 | 4.64 |

## Conclussion

Based on the results of the analysis of the multidimensional nonlinear model, the theory response items on the three logistical parameters that were applied varied in the number of items, namely 20 items, 30 items, and 40 items which had varying correlations. The factor analysis carried out for item 20 shows that the participant is higher in answering items correctly, the same is the case when it is carried out on the test length of 30. It is different from the length of the 40 item test. It can be concluded that using a sample size of 1500 with more than 40 items can differentiate the abilities of the participants. Where participants answered low on one dimension, it turned out to be able to answer correctly on another dimension.

## Limitations and Future Studies

There are many weaknesses in the factor analysis that I did. Because I still study a lot of multidimensional nonlinear models of these three logistic parameters. Another weakness is the lack of understanding of this research, making it difficult to find friends in discussions.
This study will be very interesting if further research is carried out for factor analysis that measures the correlation between the item dimensions deeper and further deepens the MIRT noncompensated model.

## Acknowledgement

Thanks, to my promotor and co-promotor in Postgraduate UNJ who have supported me in carrying out this study. I am very overwhelmed in finding discussion partners. This is due to the lack of Indonesians researching the nonlinear MIRT model.

## References

[1] Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education.

[2] Amemiya, Y. (1993a). Instrumental variable estimation for nonlinear factor analysis. In Multivariate Analysis; Future Directions 2. (C.M. Cuadras, C.R. Rao, Ed.), pp. 113-129,

North-Holland Series in Statistics and Probability, Amsterdam.

[3] Amemiya, Y. (1993b). On nonlinear factor analysis. To appear in *Proceedings of ^5^4, Social Statistics Section.* Amsterdam.

[4] Amemiya, Y. and Fuller, W. A. (1988). Estimation for the nonlinear functional relationship. *Annals of Statistics.*

[5] Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement.*

[6] Bock, R. D, Gibbons, R, and Muraki, E (1988). *Full information item factor analysis*. *Applied Psychological Measurement*.

[7] Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. Journal of Educational and Behavioral Statistics.

[8] Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika.

[9] Bock, R. D., & Lieberman, M. (1970). Fitting a response model to n dichotomously scored items. Psychometrika.

[10] Cheng, Y.-Y., Wang, W.-C., & Ho, Y.-H. (2009). Multidimensional Rasch Analysis of a Psychological Test With Multiple Subtests: A Statistical Solution for the Bandwidth--Fidelity Dilemma. *Educational and Psychological Measurement.*

[11] Christoffersson, A (1975). *Factor analysis of dichotomyzed variables*. Psychometrika.

[12] Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*.

[13] DeMars, C. E. (2006). *Scoring Subscales Using Multidimensional Item Response Theory Models*. Paper dipresentasikan pada the Annual Meeting of the American Psychological Association, Washington, DC.

[14] Etezadi-Amoli, J. and McDonald, R. P. (1983). A second generation nonlinear factor analysis. *Psychometrika*

[15] Gibbons, RD, Bock, RD, Hedecker, D, Weiss, DJ, Segawa, E, Bhaumik, DK, Kupfer, DJ, Frank, E, Grochocinski, VJ, and Stover, A (2007). *Full-information item bi-factor analysis of graded response data. Applied Psychological Measurement*.

[16] Hambleton, R.K. & Swaminathan H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Inc.

[17] Ilker Yalcin. (1995). *Nonlinear Factor Analysis*. Retrospective Theses and Dissertations. USA: IOWA State University.

[18] Jamshid Etezadi Amoli dan Roderick P. McDonald (1983). A Second Generation Nonlinier Factor Analysis. Macquarie University: Psychometric Society.

[19] Kahraman, N., & Thompson, T. (2011). Relating Unidimensional IRT Parameters to a Multidimensional Response Space: A Review of Two Alternative Projection IRT Models for Scoring Subscales. *Journal of Educational Measurement.*

[20] Knol, D. L., & Berger, M. P. F. (1991). *Empirical comparison between factor analysis and multidimensional item response models*. Multivariate Behavioral Research.

[21] Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

[22] Mark D. Reckase dan Robert L. McKinley, (1991). *The Discriminating Power of Items That Measure More Than One Dimension*. Applied Psychological Measurement. Sage: http://www.sagepublications.com.

[23] McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement.*

[24] McDonald, R.P (2000). *A basis for multidimensional item response theory*. *Applied Psychological Measurement*.

[25] McKinley, R. L dan Reckase, M. D (1983a). An Application of a Multidimensional

Extension of The Two-Parameter Logistic Latent Trait Model (ONR-83-3) (ERIC Document Reproduction Service No. ED 240 168).