Multidimensional Item Response Utilization for Validating Mathematics National Examination in Indonesia

J Arlinwibowo, I Achyani, and G Kurniadi
Universitas Muhammadiyah Kudus, Jl. Ganesha Raya No.1, Purwosari, Kec. Kota Kudus, Kabupaten Kudus, Jawa Tengah 59316

Abstract. The study purposes to analyze mathematics national examination (UN) items, which are guessed to be multidimensional, with MIRT to see test dimensions and item parameters. The analysis is expected to conclude quality of item and test more precisely. This study will examine the response patterns of 3299 DIY students who work on 40 multiple choice mathematics questions package 1 in UN in 2015. The research process begins with Principal Component Analysis (varimax rotation) analysis, analysis fit model of test with signed chi-squared, estimating parameters based on on the dimention of factors according to the results of the PCA, and the conclusion. The results of the analysis show that the Mathematics Examination Package 1 is divided into 3 factors, there are several items that measure more than one ability, and there are 1 items that must be removed because the difficulty level (MDIFF) is in the very easy category.

1. Introduction
Assessment is one important component in education. According to [3], the function of assessment is to determine achievement and motivate students to learn. [14] stated that through the results of the assessment, various parties (executor or policy maker) can take appropriate policies (learning processes or learning strategies). Thus, the results of the assessment must be able to show the true ability of students so that the follow-up effect becomes on target.

Assessment is needed at the micro or macro level, school level to a national level. For macro assessment, Indonesia has a summative assessment strategy in the form of a national exam. The implementation of the national exam is considered very urgent to map the learning outcomes of Indonesian students which are spread widely ranging from Sabang to Merauke. However, guaranteeing the quality of the results of the National Examination is not an easy, one of the most difficult is guaranteeing the quality of the items in the examination.

The process of developing questions is not easy and very expensive. It is very difficult to be able to produce a items or test packages with the best quality every year. Thus, collecting questions that were created in previous years is one of the effective solutions to create a items bank that provides items having quality. Therefore analyzing the previous questions based on the UN results data becomes very important for sorting, selecting, and concluding quality items.

Classical analysis techniques have been used since the 20th century [2]. Classical theory assumes that each individual has an true score that would be obtained if there were no errors in the measurement [11]. The paradigm of classical test theory is that the observed test score (TO) consists of an true score (T) and an error score (E) where the true score and the error score are mutually independent. [16] and [12] are illustrated theory of classical test in the formula TO = T + E. The focus
of classical theory analysis is on total test scores; frequency of correct responses (to indicate difficulty of questions); frequency of responses; test reliability and item-total correlation [9]. Although this statistic has been widely used, one limitation is the statistics that describe items depending on the sample [6].

Modern test theory or known as the item response theory developed to overcome the limitations that exist in classical test theory. The benefit of item response theory is that identification of test reliability and measurement result errors is seen through the item information function that is calculated for each item [10]. This mechanism provides a solid basis for selecting items in constructing a test [11]. Item response theory is based on invariant item parameters related to characteristic that the sample and measurement results is independent each other. In item response theory, item parameters do not depend on the sample in the sense that it does not change when it is counted on samples of other groups with different abilities. Thus, it means that a uniform measurement scale can be provided for use in various groups including when the characteristics are different [1]. Item response theory is suitable for analyzing item parameters with a large number of test participants [15], namely 300 participants for 1PL or Rasch Model [5], 500 participants for 2PL [17], and 1000 participants for 3PL [13].

In general, the analysis conducted is based on unidimensional item response theory so according to [7] a test must only measure one ability. However, there is a condition that is relatively difficult to fulfill namely that empirical data must show unidimension. But in fact, a test is very complex and tends to measure various things so that many empirical data events show that a test is grouped in several factors (multidimensional).

[4] states that to analyze test that are proven to measure more than one ability can use multidimensional item response theory (MIRT). Rackase (2009) said that the MIRT logistics model is as follows:

\[ P(X = 1|\theta, a, d) = \frac{e^{a_1 \theta_1 + d}}{1 + e^{a_1 \theta_1 + d}}, \]

Where \( \theta = \theta_1, \theta_2, \theta_3, ..., \theta_k \) is a vector of student ability and \( a = a_1, a_2, a_3, ..., a_k \) is a vector of item discrimination and \( d \) is the item's difficulty level. The discrimination index of items for a combination of various discrimination and difficulty level in each dimension of ability can be concluded with the following formula:

\[ MDISC = \sqrt{\sum_{j=1}^{k} a_j^2}, \quad MDIFF = \frac{-d}{MDISC}. \]

Thus it will be carried out an analysis of mathematics national examination, which are hypothesized that the result data of mathematics national examination grouping in various factors (multidimensional), with MIRT to see the dimensions and item parameters of the test. The analysis process is expected to produce a conclusion of item and test quality more precisely.

2. Method
This study involved students’ responses to the 2015 National Mathematics Examination in Yogyakarta. The number of students involved in the test is 3299 students. The National Examination Package analyzed is package 1 which consists of 40 multiple choice items. The first process carried out was Principal Component Analysis (PCA) with varimax rotation to show the distribution of factors from empirical data collected. The second step is to test the fit model by utilizing the packages mirt in the R software based on the signed chi-squared test. From the results of the fit test, the p-value and RMSEA will be provided with the fit model data if p-value > 0.05 and PMSEA <0.08. The third step is to estimate parameters based on the division of factors in accordance with the results of the PCA to bring up MDISC and MDIFF. The last step is the conclusion of the analysis results.
3. Result

3.1 Determination of Dimension Based on Principal Component Analysis (PCA)

In analyzing data of test using item response theory, the first step to do is to know the dimensions of the empirical data that is obtained. The testing process is carried out by exploratory factor analysis based on PCA. The initial assumptions that must be fulfilled in the exploratory factor analysis are the KMO and Bartlett tests.

Based on the KMO and Bartlett's Test, it can be said that the analyzed data set has sufficient sample indicated by $KMO = 0.986 > 0.5$. In addition, the significance of the Bartlett's test shows that the Hypothesis Zero correlation matrix which is an identity matrix is rejected ($\text{sig.} = 0.000 < 0.05$) so that the data forms a correlation matrix that shows there is a close relationship between variables. Then, based on analysis of all anti images correlation data then the measures of adequate sampling (MSA) have showed values above 0.5 so that all data have been adequate to be proceeded in the factor analysis. In the analysis of the factors generated grouping data as follows.

| Component | Initial Eigenvalues | Total | % of Variance | Cumulative % |
|-----------|---------------------|-------|---------------|--------------|
| 1         | 13.518              | 33.796| 33.796        |              |
| 2         | 1.581               | 3.952 | 37.748        |              |
| 3         | 1.014               | 2.534 | 40.281        |              |
| 4         | 0.932               | 2.329 | 42.611        |              |

Factor analysis using the Principal Component Analysis (PCA) method reduces or minimizes the number of observed variables so that they become the main components that the number are smaller than the number of items. The main component, which then becomes a factor, is determined by selecting components that have eigenvalues of more than 1. The value is a reference to calculate the grouping factors and their effects are shown from the cumulative variant (Guttman, 1954; Kaiser, 1960). Based on the analysis, empirical data shows that the items of mathematics national examination items are divided into 3 factors. By using varimax rotation, 40 items are grouped into 3 factors as follows.

| Item  | Component | Item  | Component | Item  | Component |
|-------|-----------|-------|-----------|-------|-----------|
| B1*   | 0.33      | 0.60  | 0.41      | B11   | 0.96      | 0.20      | B21*   | 0.59      | 0.51      | B31*   | 0.38      | 0.42      |
| B2*   | 0.59      | 0.48  | 0.58      | B12   | 0.55      | 0.20      | B22    | 0.53      | 0.51      | B32    | 0.30      | 0.49      |
| B3    | 0.61      | 0.50  | 0.50      | B13   | 0.45      | 0.51      | B23    | 0.63      | 0.51      | B33*   | 0.38      | 0.46      |
| B4*   | 0.53      | 0.37  | 0.45      | B14*  | 0.51      | 0.51      | B24*   | 0.36      | 0.61      | B34*   | 0.38      | 0.46      |
| B5*   | 0.51      | 0.45  | 0.33      | B15*  | 0.37      | 0.37      | B25    | 0.58      | 0.61      | B35    | 0.62      |
| B6*   | 0.57      | 0.40  | 0.47      | B16*  | 0.36      | 0.36      | B26    | 0.52      | 0.61      | B36    | 0.40      |
| B7    | 0.56      | 0.56  | 0.56      | B17   | 0.56      | 0.56      | B27    | 0.61      | 0.56      | B37*   | 0.49      | 0.39      |
| B8*   | 0.58      | 0.41  | 0.36      | B18   | 0.36      | 0.36      | B28    | 0.64      | 0.64      | B38*   | 0.52      | 0.44      |
| B9*   | 0.55      | 0.31  | 0.38      | B19   | 0.38      | 0.38      | B29    | 0.45      | 0.45      | B39*   | 0.46      | 0.54      |
| B10*  | 0.44      | 0.39  | 0.67      | B20   | 0.67      | 0.39      | B30*   | 0.41      | 0.39      | B40    | 0.56      |

* Items written in bold are items that measure more than one factor (dimension)
Based on the factor analysis, it can be concluded that the data are not unidimensional so that it cannot be analyzed by unidimensional item response theory. Therefore the data will be analyzed using multidimensional item response theory, in this case it contains 3 dimensions.

### 3.2 Analysis with Multidimensional IRT

The analysis begins by testing the fit of the data with the item parameter estimation model. Match testing uses the "mirt" package which produces an RMSEA and p-value output to indicate a fit. Identification of fit model from national examination data is done by combining RMSEA and p-values. Based on condition that the model is suitable if the RMSEA <0.08 and p-value> 0.05, then the estimated item parameter model that contains the most fit items is 2 logistic parameters (2PL). Therefore, the analysis used will be based on the 2PL model. The results of the parameter estimation of national exam items with 2PL method based on multidimensional item response theory are as follows.

| Item | a1  | a2  | a3  | d  | MDISC | MDIFF | Item | a1  | a2  | a3  | d  | MDISC | MDIFF |
|------|-----|-----|-----|----|-------|-------|------|-----|-----|-----|----|-------|-------|
| B1   | 1.21| 1.94| 0   | 3.36| 2.28  | -1.47 | B21  | 2.11| 1.85| 0   | 3.4 | 2.81  | -1.21 |
| B2   | 1.88| 1.45| 0   | 2.68| 2.37  | -1.13 | B22  | 1.51| 0   | 0   | 1.43| 1.51  | -0.95 |
| B3   | 0   | 3.28| 0   | 4.61| 3.28  | -1.41 | B23  | 1.77| 0   | 0   | 0.17| 1.77  | -0.1  |
| B4   | 1.27| 1.04| 0   | 1.75| 1.64  | -1.07 | B24  | 1.28| 2.18| 0   | 3.52| 2.53  | -1.39 |
| B5   | 1.43| 1.24| 0   | 2.26| 1.89  | -1.19 | B25  | 1.34| 0   | 0   | 0.56| 1.34  | -0.42 |
| B6   | 1.69| 1.03| 0   | 2.04| 1.98  | -1.03 | B26  | 2.48| 0   | 0   | 2.3  | 2.48  | -1.21 |
| B7   | 0   | 2.49| 0   | 3.18| 2.49  | -1.28 | B27  | 3   | 0   | 0   | 3.47| 3.3   | -1.05 |
| B8   | 1.67| 1.02| 0   | 1.92| 1.96  | -0.98 | B28  | 1.6 | 0   | 0   | 0.32| 1.6   | -0.2  |
| B9   | 1.38| 0   | 0.35| 0.08| 1.43  | -0.06 | B29  | 1.26| 0   | 0   | 0.72| 1.26  | -0.57 |
| B10  | 1.33| 1.12| 0   | 2.5 | 1.74  | -1.44 | B30  | 1.19| 0.97| 0   | 2.1 | 1.53  | -1.37 |
| B11  | 0.92| 0   | 0   | 0.73| 0.92  | -0.79 | B31  | 0.72| 1.14| 0   | 1.56| 1.35  | -1.16 |
| B12  | 1.68| 0   | 0   | 1.1 | 1.68  | -0.66 | B32  | 0   | 0   | 0.43| 0.92| 0.43  | -2.16 |
| B13  | 0   | 2.69| 0   | 3.19| 2.69  | -1.19 | B33  | 1.06| 0   | 1.22| 1.54| 1.62  | -0.95 |
| B14  | 1.12| 1.61| 0   | 2.05| 1.96  | -1.04 | B34  | 2.95| 0   | 0   | 2.64| 2.95  | -0.89 |
| B15  | 0.72| 1.07| 0   | 1.79| 1.29  | -1.39 | B35  | 1.77| 0   | 0   | 1.33| 1.77  | -0.76 |
| B16  | 1.37| 1.09| 0   | 2.15| 1.75  | -1.23 | B36  | 0.91| 0   | 0   | 0.415| 0.91  | -0.46 |
| B17  | 1.51| 0   | 0   | 0.88| 1.51  | -0.58 | B37  | 1.87| 0.16| 3.059| 1.88| 1.63  |
| B18  | 0   | 1.41| 0   | 1.08| 1.41  | -0.77 | B38  | 1.48| 1.16| 0   | 1.644| 1.88  | -0.87 |
| B19  | 0   | 1.76| 0   | 2.78| 1.76  | -1.58 | B39  | 1.01| 1.87| 0   | 2.297| 2.12  | -1.08 |
| B20  | 1.95| 0   | 0   | 0.64| 1.95  | -0.33 | B40  | 1.7 | 0   | 0   | 1.014| 1.7   | -0.59 |

Discriminant is often also referred to as slope. A positive slope indicates that the possibility of a correct response from students who have good abilities is higher than students whose abilities are lacking, while negative slope illustrates the opposite trend. For further analysis, a combination of discriminant and problem difficulty is presented in the (Multidimension Discriminant) MDISC and (Multidimension Difficulties) MDIFF. According to Baker (2001) and [8], the combination of distinguishing and difficulty items can be classified as follows:
All items have a good discrimination that is at least has a moderate category of discrimination. Regarding the difficulty of the item, it can be seen that 10 items can be ranked medium and 29 included in the easy classification, and there is one item that is classified very easily. The following bar chart provides an overview of item discrimination and the difficulty of 40 items.

![MDISC and MDIFF distributions](image)

### Figure 1. MDISC distribution (left) and MDIFF distribution (right)

#### Table 4. MDISC Classification (left) and MDIFF Classification (right)

| Criteria         | Description | Criteria         | Description |
|------------------|-------------|------------------|-------------|
| $MDISC \geq 1.7$ | Very high   | $MDIFF \geq 2$  | Very hard   |
| $1.35 \leq MDISC < 1.7$ | High          | $0.5 \leq MDIFF < 2$ | Hard        |
| $0.65 \leq MDISC < 1.35$ | Moderator     | $-0.5 \leq MDIFF < 0.5$ | Moderator  |
| $0.35 \leq MDISC < 0.65$ | Low           | $-2 \leq MDIFF < 0.5$ | Easy        |
| $MDISC < 0.35$  | Very low     | $MDIFF < -0.5$  | Very easy   |

4. **Discussion**

The results of empirical data analysis indicate that there is only 1 item, which is item 32nd on the Mathematics National Examination (package 1) in 2015 which must be revised because it is included in the very easy category according to the multidimensional IRT analysis. Here is the item 32nd in the examination package.

If we look at the 32nd problem, we will find the fact that basically the problem is not a simple problem to solve. The reason is that there are many combinations of steps to solve the problem, namely (1) knowing the structure of pyramid (2) has unit conversion ability, and (3) understand the concept of rounding. This problem should not be a memorization problem that is very easy to do. However, why is this problem has a very easy category? The hypothesis of the researcher, the cause of the item in the easy category is (1) square pyramid often found in the exercise questions, (2) square limas contains the concept of pythagoras and in Indonesia the concept of examination is not allowed to involve calculation aids (calculator device or software calculator in other device) so the combination
of numbers is very limited. The impact is that there is the possibility of problems with such a combination of numbers that are easily found in various literatures. Thus, the problem that should be reasoning problem transformed into memorizing problem.

However, more detailed casuistic research is still needed to evaluate the problem so that it is found as to why it is included in the very easy category and must be removed from the packages. The results of the analysis can be used to fix the item.

5. Conclusion

The National Mathematics Examination packages 1 in 2015, which consists of 40 multiple choice questions, is divided into 3 factors so that the analysis of response theory items must be done multidimensionally. There are several overlapping items which can be seen in table 3 so that one item does not only measure one ability. MIRT analysis shows that there is one item that must be remove because the level of difficulty (MDIFF) is included in the category of very easy, the other items are in good condition so it is feasible to enter the question bank.

6. References

[1] Anastasi, A. and Urbina, S., Psychological testing. New York, NY: Prentice Hall, 2002.
[2] Bechger, T. M., Maris. G., Verstralen, H. H. F. M., and Beguin, A. A., “Using classical test theory in combination with item response theory”, Applied Psychological Measurement, 27(5), pp. 319-334, 2003.
[3] Ebel, R. L. and Frisbie, D. A., Educational Measurement. New Delhi: Prentice Hall of India, 1991.
[4] Embretson, S. E. and Reise, S. P., Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum, 2000.
[5] Guyer, R. and Thompson, N. A., User’s manual for Xcalibre 4.1. St. Paul, MN: Assessment Systems Corporation, 2011.
[6] Hambleton, R. K., “Emergence of item response modeling in instrument development and data analysis”, Medical Care, 38, pp. 60-65, 2000.
[7] Hambleton, R. K. and Swaminathan, H., Item Response Theory: Principles and Applications. USA: Kluwer-Nijhoff Publishing, 1985.
[8] Hasmy, A., “Compare unidimensional & multidimensional Rasch model for test with multidimensional construct and items local dependence”, Journal of Education and Learning, 8(3), pp.187-194, 2014.
[9] Impara, J. C. and Plake, B. S., “Teachers’ ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method”, Journal of Educational Measurement, 35, pp. 69-81, 1998.
[10] Lord, F. M., Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.
[11] Magno, C., “Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data”, The International Journal of Educational and Psychological Assessment, 1(1), pp. 1-11, 2019.
[12] Novick, M. R. “The axioms and principal results of classical test theory”, Journal of mathematical psychology, 3, pp. 1-18, 1966.
[13] Patsula, L. N. and Gessaroli, M. E., A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes The annual meeting of the National Council on Measurement in Education, Proceedings April 1995, San Francisco, CA.
[14] Retnawati, H., Kartowagiran, B., Arlinwibowo, J. and Sulistyani, E., “Why are the Mathematics
National Examination Items Difficult and What Is Teachers’ Strategy to Overcome It?"  
*International Journal of Instruction*, 10(3), pp. 257-276, 2017.

[15] Sahin, A. and Anil, D., “The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory”, *Educational Science: Theory & Practice*, 17(1), pp. 321-335, 2017.

[16] Spearman, C., “The proof and measurement of association between two things”, *American Journal of Psychology*, 15, pp. 72-101, 1904.

[17] Stone, C. A., “Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of Multilog”, *Applied Psychological Measurement*, 16(1), pp. 1-16, 1992.