How many dimensions and items were enough for mathematics test in national examination? (Application of Multidimensional Logistic Model in Item Response Theory)

H Retnawati1, K Kumaidi2, G K Kassymova3, S Mam4 and O Ndayizeye5

1 Universitas Negeri Yogyakarta, Indonesia

2 Universitas Muhammadiyah Surakarta, Indonesia

3 Abai Kazakh National Pedagogical University and Satbayev University, Almaty, the Republic of Kazakhstan

4 Faculty of Education, Royal University of Phnom Penh, Cambodia

5 Department of Languages and Social Sciences, Higher Teacher-Training School of Burundi, (Ecole Normale Supérieure (ENS) du Burundi), Burundi

E-mail: heri\_retnawati@uny.ac.id

**Abstract**. This research study aims to find out: (1) the number of dimensions in the mathematics test, (2) the influence of the test length, (3) the number of testees, and (4) the number of dimensions to the test’s relative efficiency estimated by the unidimensional and multidimensional item response theory. This simulation study used the response of testees to the mathematics test of National Examination as the data model. The focus variables in this study are the test length, the number of testees, and the number of dimensions, with 40 replications in the data generating in every case. The information function value of the data was estimated applying the item response theories, of which the results were then to be compared. The results of research indicate that (1) the test can be regarded as the unidimensional test measuring the general mathematical ability, although still measuring the specific dimensions, they are spatial ability (in the 2 dimensional model), and the spatial and numeric abilities (in the 3 dimensional model); (2) in applying the item response theory, there is a tendency that the longer the item pool is, the higher the relative efficiency to a 30-item pool is, in unidimensional, bidimensional, as well as three-dimensional models; (3) the number of testees doesn’t influence the estimate of relative efficiency in the item response theory approach, (4) the more the dimension of item pool is, the higher its relative efficiency is if compared with unidimensional model.

1. Introduction

In the item response theory, the role of test reliability coefficient is replaced by the information function value of the test. The ratio of information function values of two tests is called relative efficiency. The value of relative efficiency will depend on the information function value of the two tests being compared. This comparison can be done after previously estimating the information function value in the item response theory.

So far, the value of information function are estimated by using a test assumption measuring one dominant dimension, without emphasis to the other specific dimensions which are also measured in the test. In the reality, most educational and psychological tests in several levels are multidimensional [1,2]. In line with these opinions, the results of the study indicate that most of the tests are mostly unidimensional structures, meanwhile, it was proven that as the complexity of the contents increase, the complexity of the dimensional structure tests increase as well [3]. Considering this fact, it is necessary to study the influence of the number of dimensions measured in a test to the estimation of the information function value in item response theory and the relative efficiency using item response theory.

Concerning with the test’s relative efficiency which is estimated using the ratio of the information function value, is affected by the stability of item parameter estimation. On the other hand, the stability of item parameter is affected by the number of examinees in the test [4], and the length of test [5]. Then, another study is also required to investigate the influence of test length and number of examinees to the test’s relative efficiency, and it is necessary to compare the relative efficiency estimation based on the item response theory.

Based on the background above, it is necessary to conduct a simulation study on the relative efficiency of test, which is a development of approaches using the item response theory. The variables in this study are the length of the test, the number of examinees, and the number of dimensions measured towards the estimation of test’s relative efficiency. In order to gain the data as in the real-life situation, this study needs parameter models to simulate or generate the data.

## The unidimensional item response theory

In an item analysis using the item response theory there are common assumptions that must be fulfilled in the analysis, i.e. the local independency and unidimensionality. In this theory, the probabilistic approach is used to express the relationship between the examinee’s ability and the expectation to answer correctly. This relationship is expressed using a logistic model with the parameter of difficulty index, item discriminating index, and pseudo guessing index.

In the three-parameter logistic model the relationship can be expressed as follows [6, p.17], [7, p.49]:

|  |  |
| --- | --- |
|  *= ci + (1-ci)*  | (1) |

where : ability level of the examinee, : the probability of the examinees with ability  can answer the *i*th item correctly, : discriminating index, : difficulty index for the *i*th item, : pseudo-guessing index for the *i*th item, : a natural number whose value approaches 2.78, *n*: the number of items in the test, and : the scaling factor whose value is 1.7.

Item information function is a method used to explain the power of an item in a set of test which states the power or the role of the item in reveal the latent trait measured with the test. Mathematically, the item information function is defined as follows:

|  |  |
| --- | --- |
| *Ii (θ)* =  | (2) |

where *i* = 1, 2, 3, …, *n*, *Ii(θ)*: information function of the *i*th item, *Pi(θ)*: the chance for an examinee with the ability *θ* to answer the ith item correctly, *P'i(θ)*: derivation of function *Pi(θ)* towards *θ*, *Qi(θ)*: the chance for an examinee with the ability *θ* to answer the *i*th item incorrectly.

Test information function is the amount of item information functions within the test [7, p.94]. Concerning with this, the value of test information function will be high if the items of the test also have high item information functions. The test information function *(I(θ))* can mathematically be defined as follows:

|  |  |
| --- | --- |
| *I (θ) =*  | (3) |

## The multidimensional item response theory

In the multidimensional item response theory (MIRT) there are two models, i.e. compensatory and non­compensatory. According to Ansley and Forsyth as cited in [8], the compensatory model allows a high ability in one of the dimensions to have the compensation of low ability in another dimension in relation of the probability to answer correctly. On the contrary, the noncompensatory model does not allow a high ability in one of the dimensions to have the compensation of low ability in another dimension. For the compensatory model in the case of two-dimensional item, an examinee with a very low ability in one dimension and very high ability in another dimension can answer the test item correctly. This research focused only on the MIRT in the compensatory model.

The linear logistic MIRT in the compensatory model can be written as follows:

|  |  |
| --- | --- |
| *Pi (****θ****j) = ci + (1-ci )*  | (4) |

where *fijm =* ***a****jm’* ***θ****im,* ci is the *pseudo-guessing* parameter for the *i*th item, ***a***jm: discriminating parameter for the *i*th item in the *m*th dimension, di: parameter of difficulty level for the *i*th item, and ***θ****jm*: the *m*th element of the ability vector of the *j*th examinee *(****θ****j)*.

Item information function in the multidimensional item response theory is expressed as:

|  |  |
| --- | --- |
| Iiα (θ) =  | (5) |

where Iiα (θ): information presented by the *i*th to the direction of α in the room, and ∇α = is the definitive operator for the derivations with the direction of α.

Bryant [9] presents this directed item information function in more details. The directed information function is expressed as:

|  |  |
| --- | --- |
| Ii(**θj) =** D2 **(ai’ ui)2** Qi(**θj) {** Pi(**θj)** [1+Exp(-L)]2}-1 | (6) |

where L= D **(ai’θj +** di**)**.

The ability scale which maximizes the information function value is:

|  |  |
| --- | --- |
| **θmax = ui** [ln{.5 [1 +(8ci +1)1/2]}(D“**a**i“)-1−di (“**a**i“)-1] | (7) |

## The relative efficiency of test

De Gruijter and Van der Kamp [9, p.118] state that the value of item information function and also the value of test information function depend on the latent trait or ability. The information function (I) value is invariant, so that the ratio of the information value of two items will also be invariant. The ratio of information value of two items can be written as:

|  |  |
| --- | --- |
|  | (8) |

for all transformations from θ\* to θ. The invariant feature of the ratio of test information function value is used to know the relative efficiency of the tests. Relative efficiency of two tests, according to McDonald [10, p.279] can be defined as the ratio of their error variants or, equivalently, the ratio of their information function values. These values can be compared if the two tests measure the same attributes. The same thing is also proposed by Lord [11, p.83] and Stocking [13] as well, but with different symbols.

## The dimensionality of the data

The analysis exploratory factor, which is used in this study, is a technique to detect and access the latent source of variant or covariant in a measurement [14]. An exploratory factor analysis is exploring the empirical data to find out and detect the characteristics and relationship between variables without determining the model of the data. In other words, it finds out the number of factors based on empirical data. In this analysis, the researcher does not have a priorytheory to set up the hypothesis [15]. The relationship between the item and dimension is expressed in full-model. It is this model that will be used this study, to find out how many factors are measured, which will be called dimensions.

To know the number of factors, the Eigen value must be kept greater than one. This Eigen value can be obtained in many ways, the easiest of which is by scree plot. It is further tested whether factor subtraction or addition to the previous factor is significant, by using the variance of χ2 values which are obtained by placing the k factor and the *k*+1 factor [16].

1. Method

This study consists of two parts which are related to each other, i.e. the preliminary research using the real data and the research using the simulation data. The preliminary research was done to achieve the first and second objectives of the study, i.e. to find out the number of ability dimensions and to know the characteristics of the test. These characteristics were then used to simulate the data. In relation with that, this study made use the data set from the National Examination of Mathematics for the level of Junior High School in Yogyakarta, a special province of Indonesia, administered in 2006. The data of examinees’ responses to the Mathematics test of National Examination were documentation data which were quoted to be analyzed.

Asimulation study was used to know the influence of test length, the influence of sample size of the examinees, and the influence of the number of dimensions to the test’s relative efficiency using the unidimensional and multidimensional item response theory approaches. To answer this problem, a simulation study was conducted using the estimated data brought up by setting up the desired condition, such as the situation in the field, in the form of model data.

Concerning with the first and second problems, this study also used the real data, i.e. the mathematics the Mathematics test of National Examination. Using the available patterns of the examinees’ responses, the items’ parameter, parameter of the examinees, and dimensions in the test could be estimated. These parameters were then used as models to bring up the data, so that the simulation data are similar to the data obtained in the real situation.

In accordance with the objectives of this research, the simulation study will reveal the effect of the test length, the number of test, and the number of abilities or dimensions to its relative efficiency. The index of relative efficiency is measured by using the unidimensional and multidimensional item response theory. The length of tests used in this study is 15, 20, 25, and 30 items. Tests with the length of 15 and 20 items represent the short test, while those with 25 and 30 items represent the long test. The number of examinees is a significant variable in this study. In this case, there are 3 types of examinees to be concerned with, i.e. 500 people, 1000 people and 1500 people. Concerning with the number of dimensions, there are three types of data in use, i.e. 1 dimension (unidimensional), 2 and 3 dimensions. Based on the length of the test, number of examinees, and number of abilities, there are 4×3×3 = 36 data cases generated. There are 40 replications conducted for each case to strengthen the result of this study.

1. Results

## The results of preliminary research

The set of data used in this preliminary research was the responses of the Junior High School students in Yogyakarta Indonesia to the Mathematics National Examination. It is a documentation data taken from the Provincial Office of Education in Yogyakarta. In this preliminary study, 3,012 responses were analyzed from the examinees that represented students in 4 regencies and 1 city in Yogyakarta Special Province of Indonesia. For each regency or city, the selected schools were managed to represent those located in the downtown area, in the suburban area and in the rural area (relatively far from the city center).

To find out the factors of the mathematics test, a factor analysis was conducted. Based on this analysis, it is revealed as the empirical data that the mathematics National Examination test measured better 3 factors, 2 factors and 1 factor, consecutively. These factors are then called the dimensions. The result of this statistical test supports the result of determining factors using the scree-plot technique, which shows that the mathematics test for Junior High School’s national examination measures 2 dimensions. However, for the purpose of this study, there are three dimensions used as variables, i.e. 1, 2, and 3 dimensions.

After the number of factors had been determined, the factors were then given names. Naming the factors was based on its loading after being rotated, by paying attention to the magnitude of the factors that are more than 0.4. The naming of the factors used in the test were conducted by the researcher with the assistance of a mathematics expert, practitioners (2 teachers), an expert in mathematics education and a psychologist in a forum of Focus Group Discussion (FGD). Previously, analysis had been conducted using 2 factors, using 3 factors, and using the promax rotation as well.

Based on the FGD result, for the model with 1 factor, the factor was named general ability in mathematics. For the model with 2 factors, factor I is named general ability and the second factor is named spatial ability, while for the model with 3 factors, the first one is named general ability, the second spatial ability, and the third numerical ability. The name of each factor was based on its loading after the factor had been analyzed using the non-orthogonal rotation.

After that, with the assistance of 2 senior math teachers, a lecturer in math education, a lecturer in mathematics and a psychologist, in the FGD forum, the test items were reduced by paying close attention to maintain the representation of its loading. Item reduction was conducted in some stages. In each stage 5 items were reduced (from 30 items to 25 items to 20 items to 15 items). While doing so, the characteristics of model items were set up for the 3-parameter logistic model of unidimensional, bidimensional and 3-dimensional parameters with the test length of 30 items, 25 items, 20 items and 15 items as the models to bring up data in the simulation research.

## The result of simulation study

Kromrey et al. [17]. Then the data were analyzed. In the analysis using the item response theory approach, the items were previously analyzed using BILOG [18] to estimate the *pseudo-guessing* parameter (c parameter). This c parameter is then made as the input for the program syntax to estimate the item and ability parameters by means of TESTFACT [19]. After that, the information function value is estimated using the MATLAB. The estimation results for all cases were then averaged out on the 40th replications.

After the information function value has been estimated, the relative efficiency of item groups inter-cases can be estimated. The estimation results of relative efficiency for groups of items based on their dimensions is presented in Figure 1. The figure shows that the highest estimation results of the whole cases is in 3-dimension model compared to 1-dimension model, followed by 3-dimension model to 2-dimension model. The lowest estimation result is achieved in 2-dimension model compared to 1-dimension model. In the picture, the influence of test length and number of examinees towards the test relative efficiency estimation does not show a certain pattern. The relative efficiency on each dimension and each group of items is compared to the 30-item group. The result of 1-dimension model is presented in Figure 2, that of 2-dimension in Figure 3, and that of 3-dimension in Figure 4.

If we observe the result of 1-dimension cases, the estimation results of relative efficiency for groups of items with the length of 15, 20, 25, and 30 items compared to the 30-item group vary. There is an increase of relative efficiency value from the 15-item group compared to 30 items to the 20-item group compared to 30 items. But after that, the value decreases slightly when the 25-item group is compared to 30 items, and increases again when the efficiency 30-item group is compared to 30 items, and shows the tendency that the longer the test (more items), the higher the relative efficiency value is to the 30-item group. The result shows a tendency, an influence of the increase of test length to the value of relative efficiency compared to the 30-item group.



**Figure 1.** The graph of average estimation result of relative efficiency using item response theory approach (Note: n: number of items, N: number of examinees, and K: number of dimensions)



**Figure 2.** The graph of average estimation result of relative efficiency using item response theory approach on the simulation data to the 30-item group with 1 dimension (Note: N500 (500 examinees), N1000 (1000 examinees), N1500 (1500 examinees))



**Figure 3.** The graph of average estimation result of relative efficiency using item response theory approach on the simulation data to the 30-item group with 2 dimensions (Note: N500 (500 examinees), N1000 (1000 examinees), N1500 (1500 examinees))

The results for 2 dimensions are quite similar to those for 1 dimension. It shows that the estimation results of relative efficiency to the groups with 15, 20, 25, and 30 items compared to the 30-item group vary. There is an increase of relative efficiency value from the 15-item group compared to 30 items to the 20-item group compared to 30 items. The value decreases slightly when the 25-item group is compared to 30 items, and increases again when the 30-item group is compared to 30 items, and shows the tendency that the longer the test (more items), the higher the relative efficiency value is to the 30-item group. The result shows a tendency, an influence of the increase of test length to the value of relative efficiency compared to the 30-item group.



**Figure 4.** The graph of average estimation result of relative efficiency using item response theory approach on the simulation data to the 30-item group with 3 dimensions (Note: N500 (500 examinees), N1000 (1000 examinees), N1500 (1500 examinees)

In the 3 dimensions, the increase of relative efficiency estimation results to the group with 15, 20, 25, and 30 items compared to the 30-item group is clearer, except for that with 500 examinees (which decreases slightly when the group length is 20 items, but then increases again). Based on the graph of 3-dimension model, there is an increase from the 15-item group compared to 30 items to the 20-item group compared to 30 items, to the 25-item group compared to 30 items, and increases again when the 30-item group is compared to 30 items. The result shows a tendency, an influence of the increase of test length to the value of relative efficiency compared to the 30-item group.

1. Discussion

## Number of dimensions measured in mathematics test for national examination

Observing the result of exploratory factor analysis, we mathematics test for National Examination measured 1 dominant factor as well as other factors. If it only measured 1 dominant factor, the contribution of this main factor is only 44.29% in explaining the total variants. This contribution is relatively far from the value of 100%, the percentage usually expected by those test makers for the test to explain the various traits or abilities of the in National Examination examinees. The contribution of this main factor can be improved by adding other factors which are also measured in the mathematics test in National Examination.

Considering the Eigen value of the factor analysis, there are 4 Eigen values which are more than 1. This indicates that there are 4 factors that are possible to give considerable contribution to the total variants that can be explained. However, not all of them might be significant if included as the factors contained in the mathematics test for the in National Examination. By using the Chi-square test, it is found that the factor analysis model with 2 factors is better than the factor analysis model with only 1 factor. In the same way, model with 3 factors is better than that with 2 factors. However, model with 4 factors is not better than the model with 3 factors. It can be concluded then that there are 3 factors included in the mathematics test for in National Examination.

By adding one more factor to the initial model of analysis, we get 2-factor analysis, and there is an increase in the total variance that can be explained. The contribution of these 2 factors in explaining the variance is 52.19%. In other words, there is an increase of 7.90% for the contribution only by adding 1 more factors for the analysis. If another factor is added to the analysis to make it a 3-factor analysis, the contribution of variants measured to be 55.76% or having an increase of 3.57%. It means that the greatest variance is contributed by the first factor only, while the contribution of second and third factors altogether in explaining the variants is less than that if the first factor.

In the factor analysis, the first Eigen value is the greatest compared to the second, or the third Eigen and so forth. The magnitude of this Eigen value shows the linear dependency to the data. In the second, third, and so forth Eigen values, the magnitude is quite small compared to the first Eigen value [20]. Since the size of variants that can be explained by factors is consistent with the Eigen value, the first factor in a factor analysis gives the greatest contribution compared to the other factors.

Accordingly, there are 3 factors measured in the mathematics test for in National Examination. It means that the mathematics test measures at least three ability factors, which is called dimensions in this study. According to the factor loading after being rotated, these three factors were then given names. As suggested by the expert in mathematics, the 1-factor analysis is called general mathematical ability, the 2-factor analysis is called general and spatial abilities, and the 3-factor analysis is called general, spatial, and numerical abilities. It shows that there are several dimensions measured in the math test, or in other words, the math test can measure 1 dimension of ability, i.e. general mathematical ability, 2 dimensions of ability, i.e. general and spatial abilities, and 3 dimensions of ability, i.e. general, spatial, and numerical abilities.

The result of this study shows that the mathematics test for in National Examination measures more than 1 dimension or contains multidimensional loading. This is supported by Reckase [21], Bolt and Lall [1], Ackerman, Gierl, and Walker [2] and is in line with the study by Thurber, Shinn, and Smolkowski [22], that an achievement test measures more than one dimensions. Similarly, Kartowagiran and Retnawati [23] states that the mathematics tests for in National Examination in 2003 and 2005 measure more than one dimensions of ability.

## The influence of test length and number of examinees to the test’s relative efficiency

In the estimation of relative efficiency to the 30-item group, there is a tendency that the estimation result of relative efficiency value increases. This increase happens in the groups from 15, 20, 25, to 30 items, both in the unidimensional, 2-dimensional and 3-dimensional data. The tendency in the analysis shows that there is an increase of relative efficiency value in the groups from 15, 20, 25, to 30 items, both estimated in the unidimensional, 2-dimensional and 3-dimensional data with item response theory approach. It is so because the information function value of the groups of items in a scale of ability constitutes the sum of information function value of its items in that scale of ability. The longer a group of items compared to the information function value of the 30-item group, the greater its estimated relative efficiency value is.

In the estimation of relative efficiency using item response theory, the relative efficiency of a group of items is estimated, and then average of all participants is calculated. Since it is the average value, this relative efficiency of estimation result does not depend on the number of participants. This is why the number of participants does not influence the relative efficiency value of estimation result.

## The influence of number of dimensions to the test relative efficiency

Based on the analysis of the average relative efficiency in the simulation data of two- and thee-dimension model compared to the one-dimension model, it is found that comparing the two-dimension to one-dimension, the estimated relative efficiency value is greater than 1. The analysis of comparing the 3-dimension to 1-dimension models and the 3-dimension to 2-dimension models also give similar results.

In the simulation data with 2- and 3-dimension models compared to 1-dimension model, we find that comparing the 2-dimension model to 1-dimension model results in estimation result of relative efficiency greater than 1. Analyzing the 3-dimension model to 1-dimension model and the 3-dimension model to 2-dimension model gives even a greater value. It shows that the 2-dimension model giver better information than the 1-dimension model and 3-dimension model gives better information than the 1-dimension model. Similarly, the 3-dimension model gives better information than the 2-dimension model.

In the unidimensional model analysis, information is obtained only from one ability, while in the bidimensional and 3-dimensional analysis; the information will be better and more detailed. If the test information function value is estimated, using more than one estimation will surely results in a greater relative efficiency value.

1. Conclusion

Some conclusions drawn from this study are: (1) set of National Exam of Mathematics subject for Junior High School can be regarded unidimensional which measured the general mathematical ability of the examinees, although it also measured some other specific dimensions, i.e. the spatial ability (in 2-dimensional model) and numerical ability (in 3-dimensional model); (2) using the item-response theory approach, there was a tendency of increasing the relative efficiency value of estimation result for the test with 15 items, 20 items, 25 items, up to 30 items towards the 30-item test, either the test is analyzed using the unidimensional, 2-dimensional or 3-dimensional models; (3) the number of examinees does not influence the test relative efficiency estimation using the item-response theory; and (4) in the item response theory approach, the relative efficiency value as the estimation result tends to increase in the tests with items 1, 2, and 3 dimensions compared to the test with unidimensio­nal items.

References

1. Bolt D M and Lall V F 2003 Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo *Appl. Psychol. Meas.* **27** 395–414
2. Ackerman T A, Gierl M J and Walker C M 2003 Using multidimensional item response theory to evaluate educational and psychological tests *Educ. Meas. Issues Pract.* **22** 37–53
3. Joaristi L and Lizasoain L 2008 Estudio de la dimensionaidad empleando análisis factorial clásico y análisis factorial de información total: Análisis de pruebas matemáticas de primaria (5o y 6o cursos) y secundaria obligatoria [Study of dimensionality using classic factor analysis and full information factor analysis: analysis of primary (5th and 6th degrees) and compulsory secondary mathematics tests] *Reli. Rev. Electrónica Investig. y Evaluación Educ. [E-Journal Educ. Res. Assess. Eval.]* **14** 1–18
4. Segall D O 2001 General ability measurement: An application of multidimensional item response theory *Psychometrika* **66** 79–97
5. Cohen A S, Kane M T and Kim S-H 2001 The precision of simulation study results *Appl. Psychol. Meas.* **25** 136–45
6. Hambleton, R.K. Swaminathan H and Rogers H J 1991 *Fundamental of item response theory* (Newbury Park, CA: Sage Publication)
7. Hambleton R K and Swaminathan H 1985 *Item response theory* (Boston, MA: Kluwer)
8. Spray J A, Davey T C, Reckase M D, Ackerman T A and Carlson J E 1990 *Comparison of two logistic multidimensional item response theory models (Research report ONR90-8)* (Iowa City, IA)
9. Bryant D U Directional item information for the multidimensional three-parameter logistik model *Run. Head Multidimens. Item Inf.*
10. de Gruijter D N M and van der Kamp L T J 2005 *Statistical test theory for education and psychology* (Graduate School of Education, Universiteit Leiden, Netherlands)
11. McDonald R P 1999 *Test theory: A unified treatment* (Mahwah, NJ: Lawrence Erlbaum)
12. Lord F M 1980 *Application of item response theory to practical testing problems* (Hillsdale, NJ: Lawrence Erlbaum)
13. Stocking M L 1999 Item response theory *Advances in measurement in educational research and assessment* ed G V. Masters and J P Keeves (Amsterdam: Pergamon)
14. Jöreskog K G and Sörbom D 1993 *LISREL 8: Structural equation modeling with the SIMPLIS command language* (Hillsdale, NJ: Scientific Software International)
15. Stapleton C D 1997 *Basic concepts and procedures of confirmatory factor analysis*
16. du Toit M 2003 *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* (Lincolnwood, IL: Scientific Software International)
17. Kromrey J D, Parshall C G, Chason W M and Yi Q 1999 Generating responses based on multidimensional item response theory *SAS Users Group International 24*
18. Mislevy R J and Bock R D 1990 *BILOG 3: Item analysis and test scoring with binary logistic models* (Moorseville: Scientific Software International)
19. Wilson D, Wood R and Gibbons R D 1984 *TESTFACT: Test scoring, item statistics, and item factor analysis* (Mooresville, IN: Scientific Software International)
20. Johnson R A and Wichern D W 2002 *Applied multivariate statistical analysis* (Upper Saddle River, NJ: Prentice Hall)
21. Reckase M D 1997 A linear logistic multidimensional model for dichotomous item response data *Handbook of Modern Item Response Theory* ed W J van der Linden and R K Hambleton (New York: Springer) pp 271–86
22. Thurber R S, Shinn M R and Smolkowski K 2002 What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures *School Psych. Rev.* **31** 498–513
23. Kartowagiran B and Retnawati H 2007 *Validasi dimensi perangkat tes ujian nasional 2003-2006* (Yogyakarta)