REiD (Research and Evaluation in Education), 5(2), 2019, 169-182
Available online at: http://journal.uny.ac.id/index.php/reid

Parallel tests viewed from the arrangement of item numbers and alternative answers

[*1]Badrun Kartowagiran; [2]Djemari Mardapi; [3]Dian Normalitasari Purnama; [4]Kriswantoro
[1,2,3,4]Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
[*]Corresponding Author. E-mail: kartowagiran@uny.ac.id

Submitted: 23 February 2019 | Revised: 7 December 2019 | Accepted: 23 December 2019

Abstract

This research aims to prove that a parallel test can be constructed by randomizing the test item numbers and or alternative answers' order. This study used the experimental method with a post-test only non-equivalent control group design, involving junior high schools students in Yogyakarta City with a sample of 320 students of State Junior High School (SMPN) 5 Yogyakarta and 320 students of SMPN 8 Yogyakarta established using the stratified proportional random sampling technique. The instrument used is a mathematics test in the form of an objective test consisting of a five-question package and each package contains 40 items with four alternatives. The test package is randomized in the item numbers' order from the smallest to the largest and vice versa. The options in each item are also randomized from A to D and vice versa. Each item is analyzed using the Classical Test Theory and Item Response Theory approaches, while data analysis is done using the discrimination index with Kruskal-Wallis test technique to see the differences among the five-question packages. The study reveals that the result of item analysis using the Classical Test Theory and Item Response Theory approaches shows no significant difference in the difficulty index among Package 1 until Package 5. Nevertheless, according to the Classical Test Theory, there is a category shift of the difficulty index of Package 2 until Package 5 when compared to Package 1 – the original package – which is, in general, not a good package, because it contains too easy items.

***Keywords**: correct option placement, order of items, parallel test*

Permalink/DOI: https://doi.org/10.21831/reid.v5i2.23721

Introduction

National Examination (NE) is one of the government efforts to improve the quality of education. In addition to its function to measure and evaluate the achievement of school graduate competence in certain subjects, as well as to map out the quality of primary and secondary education, NE also functions as the motivator for related parties to work better in order to achieve a good examination result (Center for Educational Assessment, 2014, p. 1). The education system and teaching quality are two related matters. A good teaching system will result in good quality learning (Mardapi, 2014, p. 12). Furthermore, teaching quality can be seen from the result of the evaluation done by teachers or educators.

According to Law No. 14 of 2005 of Republic of Indonesia, teachers are professsional educators whose main duties are to educate, teach, guide, direct, drill, assess, and evaluate students of formal early-childhood education, primary education, and secondary education, and it can be understood that teachers' role is not only to plan and implement teaching, but also to assess or evaluate. The assessment of students' learning achievement is expected not only to find out whether or not the stated learning objectives have been achieved, but also to reveal whether the

objectives are important for students, and how the students achieve the learning objectives. Studies have shown that 87% teachers still find it difficult to perform assessment (Rusilowati in Rohmawati, 2013). The unsocialized procedures for conducting assessment becomes one of the constraints. This indicates that teachers' competence in doing good assessment still needs improvement.

One of the evaluation techniques which can be used to see students' competence is a testing technique. Miller, Linn, and Gronlund (2009, p. 28) define testing as an evaluation given to students for a certain period in their comparatively equal condition. A test usually consists of a set of questions. The aim of a test is to answer the question of how well an individual does something, in comparison to others or compared to the domain of performance work. The result of an assessment is the information about the characteristics of an individual or a group (Rasyid & Mansur, 2008, p. 11). In other words, using the assessment technique, a teacher is able to identify the characteristics of students' competence in a certain subject.

Based on the types of students' answers, tests can be classified into written tests, oral tests, and also performance tests (Sanjaya, 2010, p. 355). Written tests are further classified into essay tests and objective tests. One of the forms of objective tests is multiple-choice tests. Multiple-choice tests require students to choose a correct response out of some choices provided. In some choices, students choose the best choice from a list of alternatives. The choice of this type of tests is due to some consideration in relation to the strengths and weaknesses of multiple-choice tests (Reynolds, Livingston, & Willson, 2009). On the one hand, the strengths of multiple-choice tests include that a relatively large number of multiple-choice items can measure efficiently, objectively, and in a reliable manner. Besides, multiple-choice tests are very good at measuring lower cognitive objectives (for example knowledge, comprehension, and application) and they can minimize construction factors which are not relevant.

On the other hand, multiple-choice tests have weaknesses, including that they are relatively difficult to construct because more time has to be spent on making multiple-choice questions, they cannot measure all educational objectives (e.g. writing skill) although the item with alternative answers is very suitable for measuring higher cognitive domains (i.e. analysis, synthesis, and evaluation) and there are possibilities of random guessing. Identifying the strengths and weaknesses of multiple-choice tests becomes the consideration for choosing a certain type of tests.

The requirement for a good test is that it is valid and reliable (Azwar, 2013). Further, Azwar adds that measurement is said to be highly valid if it results in data that accurately give the description of the measured variable. Being accurate in this case means being exact and precise so that when a test results in data which are irrelevant to the goal of measurement, then it is considered as the measurement that has low validity. In addition to validity, the test reliability also needs to consider. The reliability of a test is said to be good if the test can result in scores/answers which are consistent although it is used by other examiners and at different time with the same condition. There are various ways of finding out the reliability of a test, one of them is the use of a parallel test (Azwar, 2015, p. 55).

Parallel tests are two or more tests whose purpose, difficulty index, and construction are the same but whose items are different. They are needed so that the alternative tests can be administered to different examinees or at different time, and also in the context of reliability estimation for the test given (Kronmüller et al., 2008; Werheid et al., 2002). So far, a parallel test can be constructed by changing the item number and or by changing the order of the alternative answers from one test package into various test packages. This has been an assumption for some teachers, while in reality there has not been a study showing that the strategy is correct, and that by changing the item number and or by changing the order of the alternative answers we can get parallel tests. For this reason, this research aims to find out the effect of randomizing item numbers and or the order of the alternative answers on item difficulty index.

Method

This research is an experiment applying the quantitative approach. It used the post-test only non-equivalent control group design to reveal the effect of randomizing item numbers and or the order of the alternative answers on item difficulty index. In this study, a mathematics test to be administered was randomized in terms of its item numbers and placement of the correct alternative to prove that the tests are really parallel. The sample was established using the stratified proportional random sampling technique. The test was reassembled into five test packages, administered to 320 students of State Junior High School (SMPN) 5 Yogyakarta and 320 students of SMPN 8 Yogyakarta. The result of the analysis shows the difference of the item difficulty index before and after the examinees do the test whose item numbers and alternative answer have undergone changes.

The data were collected using a test. The test used is a grade IX junior high school mathematics test consisting of 40 multiple-choice items with four alternative answers. The test was randomized in terms of the item numbers and alternative answers, resulting in five test packages as shown in Table 1.

Table 1. Test package and type of randomization

Test Package	Types of Randomization
1	Without randomization
2	Number randomization 1 -40
3	Number randomization 1-20 and 21-40
4	Number randomization 1-10. 11-20. 21-30 and 31-40
5	Alternative answer randomization

Table 1 shows the type of randomization of test packages. Package 1 is the original grade IX junior high school mathematics test whose item numbers and alternative answers did not undergo randomization. Package 2 is the test package resulted from the randomization of entire numbers 1–40, by reversing number 40 into number 1, number 39 into number 2, number 38 into number 3 and so on. Package 3 is the test package resulted from the randomization of numbers 1-20 and 21–40, by reversing number 20 into number 1, number 19 into number 2, and so on. Furthermore, number 40 is reversed into number 21, number 39 into number 22, and so on. Package 4 is the test package resulted from the randomization of numbers 1-10, 11-20, 21-30 and 31-40, by reversing number 10 into number 1, number 9 into number 2, and so on. Further, number 20 was reversed into number 11, number 19 into number 2, and so on. Number 30 was reversed into number 21, number 29 into number 22, and so is the case with numbers 31-40. Package 5 is the test package resulted from the randomization of alternative answers, where alternative d becomes alternative a, alternative c becomes alternative b, alternative b becomes alternative c, and alternative a becomes alternative d.

The result of administering the parallel tests to students was analysed to see if the randomization of item numbers and alternative answers really resulted in parallel tests. The data in the form of students' answer sheets were analysed using the Clasical Test Theory and Item Response Theory. The analysis using the Clasical Test Theory was done with the help of the *QUEST* program to see the item difficulty index, and the analysis using the Item Response Theory used the one-parameter logistic approach employing the *QUEST* program to see the item difficulty index and student competence parameter.

In addition to being analysed using the Clasical Test Theory, the data were analysed for their reliability index. The reliability index was seen based on the output of the analysis using the *QUEST* program. The analysis of the effect of the randomization of item numbers on the test's being parallel was conducted using Kruskall-Wallis analysis with the help of the SPSS Program. If the result of the analysis shows Sig < 0.05 then there is an effect of the randomization of item numbers of the parallel tests on the difficulty index of the tests.

Findings and Discussion

Findings

The findings of this study show that the instrument reliability index of the five test packages is good. Mehrens and Lehmann (1973, p. 122) write that the minimum reli-

ability index of an instrument is 0.80. Package 1, Package 2, and Package 3 have the same reliability index of 0.96, while the reliability index of Package 4 and Package 5 is 0.97. This finding shows that the instrument reliability index of the five test packages is in the reliable category. Furthermore, the five packages were analysed using the Classical Test Theory and the Item Response Theory. This analysis is conducted in order to find out the test parallelism based on the test item difficulty.

Parallel Tests Based on the Classical Test Theory Approach

Before scrutinizing whether or not Package 2, Package 3, Package 4, and Package 5 are parallel to Package 1, the test item characteristic funcion was studied based on the Classical Test Theory. According to the Classical Test Theory, whether a test is good or not depends on the value of the difficulty index, discrimination power, and the functioning of the distractors. Allen and Yen (1979) classify item difficulty indices into three cate-

Table 2. Characteristic function of test Package 1 based on the Classical Test Theory

Item Numbers	Difficulty Index	Category	Discrimination Power	Category	Distractor	Conclusion
1	0.945	Easy	-0.07	Poor	All Functioning	Poor
2	0.874	Easy	0.29	Good	All Functioning	Poor
3	0.929	Easy	0.15	Poor	All Functioning	Poor
4	0.134	Difficult	0.15	Poor	All Functioning	Poor
5	0.638	Moderate	0.35	Good	All Functioning	Good
6	0.709	Easy	0.28	Good	All Functioning	Poor
7	0.961	Easy	0.14	Poor	All Functioning	Poor
8	0.724	Easy	0.34	Good	All Functioning	Poor
9	0.937	Easy	0.27	Good	All Functioning	Poor
10	0.701	Easy	0.36	Good	All Functioning	Poor
11	0.748	Easy	0.29	Good	All Functioning	Poor
12	0.480	Moderate	0.47	Good	All Functioning	Good
13	0.827	Easy	0.16	Poor	All Functioning	Poor
14	0.953	Easy	0.22	Good	A and D not functioning	Poor
15	0.449	Moderate	0.24	Good	All Functioning	Good
16	0.740	Easy	0.74	Good	All Functioning	Poor
17	0.913	Easy	0.35	Good	All Functioning	Poor
18	0.827	Easy	0.24	Good	All Functioning	Poor
19	0.646	Moderate	0.26	Good	All Functioning	Good
20	0.748	Easy	0.05	Poor	All Functioning	Poor
21	0.961	Easy	0.11	Poor	A and D not functioning	Poor
22	0.709	Easy	0.31	Good	All Functioning	Poor
23	0.102	Difficult	0.1	Poor	All Functioning	Poor
24	0.307	Moderate	0.21	Good	All Functioning	Good
25	0.268	Difficult	0.35	Good	All Functioning	Poor
26	0.433	Moderate	0.4	Good	All Functioning	Good
27	0.764	Easy	0.34	Good	All Functioning	Poor
28	0.921	Easy	0.34	Good	All Functioning	Poor
29	0.748	Easy	0.29	Good	All Functioning	Poor
30	0.449	Moderate	0.11	Poor	All Functioning	Good
31	0.654	Moderate	0.44	Good	All Functioning	Good
32	0.898	Easy	0.15	Poor	All Functioning	Poor
33	0.535	Moderate	0.52	Good	All Functioning	Good
34	0.654	Moderate	0.35	Good	All Functioning	Good
35	0.732	Easy	0.22	Good	All Functioning	Poor
36	0.882	Easy	0.31	Good	All Functioning	Poor
37	0.591	Moderate	0.04	Poor	All Functioning	Good
38	0.346	Moderate	0.31	Good	All Functioning	Good
39	0.693	Moderate	0.29	Good	All Functioning	Good
40	0.465	Moderate	0.38	Good	All Functioning	Good

gories. An item is in the difficult category if its coefficient is <0.3; it is in the moderate category when its coefficient is between 0.3 and 0.7, and it is in the easy category when the coefficient is >0.7. A good test item has the difficulty index in the moderate category. The item discrimination power was also used as a consideration in deciding if test item is good or poor. Fernandes (1984) states that a good test item is an item with the discrimination power of >0.2. He adds that a distractor is considered functioning when it is chosen by at least 2% of the total examinees.

Table 2 shows the analysis result of the item characteristic function based on the Classical Test Theory. It shows the characteristics of the 40 test items in Test Package 1. In general, the items in Test Package 1 are in a poor category. A test item is said to be good when it meets three categories, i.e. having a moderate difficulty index and good discrimination power, and all of its distractors function well. In Test Package 1, only fourteen items (35%) are in a good category, while 26 items (65%) are in a poor category. Table 2 also shows that 23 items (57.5%) are categorized as easy.

Table 3. Difficulty index of five test packages (Classical Test Theory approach)

Item Number	Item Difficulty Index				
	Package 1	**Package 2**	**Package 3**	**Package 4**	**Package 5**
1	0.945	0.952	0.641	0.651	0.954
2	0.874	0.839	0.516	0.914	0.868
3	0.929	0.903	0.730	0.638	0.829
4	0.134	0.129	0.897	0.063	0.033
5	0.638	0.508	0.468	0.638	0.829
6	0.709	0.508	0.754	0.533	0.638
7	0.961	0.903	0.817	0.330	0.967
8	0.724	0.702	0.833	0.829	0.638
9	0.937	0.902	0.683	0.868	0.914
10	0.701	0.637	0.714	0.954	0.661
11	0.748	0.637	0.889	0.586	0.638
12	0.480	0.306	0.873	0.645	0.408
13	0.827	0.742	0.770	0.816	0.697
14	0.953	0.847	0.079	0.875	0.941
15	0.449	0.435	0.460	0.737	0.474
16	0.740	0.750	0.675	0.474	0.737
17	0.913	0.895	0.921	0.941	0.875
18	0.827	0.863	0.619	0.697	0.816
19	0.646	0.694	0.881	0.408	0.645
20	0.748	0.398	0.651	0.638	0.586
21	0.961	0.919	0.714	0.428	0.961
22	0.709	0.661	0.849	0.632	0.632
23	0.102	0.266	0.651	0.743	0.566
24	0.307	0.839	0.675	0.658	0.493
25	0.268	0.782	0.722	0.250	0.349
26	0.433	0.331	0.849	0.349	0.250
27	0.764	0.766	0.611	0.493	0.658
28	0.921	0.935	0.556	0.566	0.743
29	0.748	0.726	0.683	0.632	0.632
30	0.449	0.460	0.206	0.961	0.428
31	0.654	0.734	0.968	0.283	0.684
32	0.898	0.863	0.556	0.664	0.862
33	0.535	0.718	0.278	0.553	0.618
34	0.654	0.480	0.302	0.645	0.724
35	0.732	0.815	0.325	0.822	0.724
36	0.882	0.895	0.325	0.724	0.822
37	0.591	0.629	0.786	0.724	0.645
38	0.346	0.653	0.857	0.618	0.553
39	0.693	0.718	0.754	0.862	0.664
40	0.465	0.323	0.484	0.684	0.283
Average (*b*)	**0.675**	**0.677**	**0.651**	**0.661**	**0.668**

Based on the analysis from Test Package 1, the difficulty index of each item in Test Packages 2, 3, 4, and 5 was analyzed. The item difficulty index analysis using the Classical Test Theory was done with the QUEST program. The analysis result of the parameter of item difficulty index of each test package is shown in Table 3. It shows the difficulty index of each item in the five test packages. Package 1 is the original test package without any randomization, so Package 2, Package 3, Package 4, and Package 5 which had undergone randomization were reconstructed to their former forms with item numbers being rearranged to their original arrangement.

Table 4 shows that after the randomization of item numbers and alternative answers, the difficulty index of the five packages ranged from 0.102 to 0.968. This range is quite large, because, according to the Classical Test Theory, the difficulty index should range from 0 to 1. Further, based on the result of the analysis shown in Table 4, the characteristics of each test items in the five packages was analysed. The result of analysis of the each test item characteristics in terms of difficulty index is shown in Table 4.

Table 4 shows that all five test packages, viewed from the difficulty index, generally show that the test items are in easy and moderate categories. The test packages have undergone randomization and have been reconstructed into their former construction before randomization. It can be seen from the same proportion of the test packages, while the number of the items in the difficult category is only two or three. A deeper look into it reveals that some items have gone through changes in the category of difficulty index. For instance, Item 6 in Package 1 was categorized as an easy item, but after the randomization in Package 2, it was categorized as a moderate item. Another example is Item 25 in Package 1, categorized as a difficult item, but after the randomization of Package 2, in Package 3 it was categorized as an easy item. It shows that seen from the difficulty index category, many items change after the item numbers are randomized. The percentages of the changes or shifts in the item difficulty category is shown in Table 5.

Table 5 shows that the biggest shift in difficulty index is the shift of 24 items (60%) from Package 1 to Package 3, while the smallest shift is the shift from Package 1 to Package 2, i.e. 9 items (22.5). Based on the result of the analysis using the Classical Test Theory approach, Kruskall-Wallis analysis was conducted to see whether there was any significant difference of the item difficulty index of the randomized test packages. The summary of the result of the analysis is in Table 6.

Table 4. Characteristics of item difficulty index based on Classical Test Theory

Category	Package 1 (Item Number)	Package 2 (Item Number)	Package 3 (Item Number)	Package 4 (Item Number)	Package 5 (Item Number)
Easy	1, 2, 3, 6, 7, 8, 9, 10, 11, 13, 14, 16, 17, 18, 20, 21, 22, 27, 28, 29, 32, 35, 36	1, 2, 3, 7, 8, 9, 11, 13, 14, 16, 17, 18, 21, 24, 25, 27, 28, 29, 31, 32, 33, 35, 36, 39	3, 4, 6, 7, 8, 10. 11, 12, 13, 17, 19, 21, 22, 25, 26, 31, 37, 38, 39	2, 8, 9, 10, 13, 14, 15, 17, 23, 20, 35, 39	1, 2, 3, 5, 7, 9, 14, 16, 17, 18, 21, 28, 34, 35, 32, 34, 35, 36
%	57.5%	60%	47.5%	30%	45%
Moderate	5, 12, 15, 19, 24, 26, 30, 31, 33, 34, 37, 38, 39, 40	5, 6, 10, 12, 15, 19, 20, 22, 26, 30, 34, 37, 38, 40	1, 2, 5, 9, 15, 16, 18, 20, 23, 24, 27, 28, 29, 32, 34, 35, 36, 40	1, 3, 5, 6, 7, 11, 12, 16, 18, 19, 20, 21, 22, 24, 26, 27, 28 29, 32, 33, 34, 36, 37, 38, 40	6, 8, 10, 11, 12, 13, 15, 19, 20, 22, 23, 24, 25, 27, 29, 30, 31, 33, 37, 38, 39
%	37.5%	35%	45%	62.5%	52.5%
Difficult	4, 23, 25	4, 23,	14, 30, 33	4, 25, 31	4, 26, 40
%	7.5%	5%	7.5%	7.5%	7.5%

Table 5. Category shift of item difficulty index of five test packages

Packages 1-2	Packages 1-3	Packages 1-4	Packages 1-5
9 items (22.5%)	24 items (60%)	20 items (50%)	15 items (37.5%)

Table 6 shows that the value of Asymp, Sig in all items whose discrimination power is tested among Package 1, Package 2, Package 3, Package 4, and Package 5 is above 0.05. It means that there is no difference in difficulty index of the items in all five test packages, so there is no effect item number randomization on the item difficulty index. After the effect of item number randomization was scrutinized, the effect of the randomization on discrimination index was analysed. The percentages of the good and poor discrimination index is shown in Table 7.

Table 7 shows that the discrimination index of Test Packages 1, 2, 3, 4, and 5 is in a good category (> 60%). Based on the analysis of Test Package 1, after the randomization of Test Packages 2, 3, 4, and 5, there is a shift in the good discrimination index. However, a closer look reveals that the shift is not big enough, occurring to two to four items only.

Parallel Tests Based on Analysis Using Item Response Theory Approach

Before scrutinizing whether Package 2, Package 3, Package 4 and Package 5 are parallel to Package 1 or not, the researchers need to describe the assumption test of the Item Response Theory (IRT), which is the unidimension assumption test (Naga, 1992). The requirement for unidimension is aimed at sustaining invariance in IRT. If a test item measures more than one dimension, then the answer to the item is a combination of different competencies of the examinees. Thus, the contribution of each competency to the answer is unknown.

Unidimension assumption testing is carried out to reveal whether a test measures one trait. The unidimension assumption is tested by the factor analysis and its empirical result. The KMO-MSA value is sufficient if it is above 0.5 (Field, 2009). By looking at the first eigenvalue contribution to test variance, according to Reckase (1979), the formation of eigenvalue factor has to have a value above 1. In the factor analysis, the first eigenvalue has to have the biggest value (dominant) compared to the second, third, and so forth eigenvalues. The result of the analysis of unidimension assumption testing is shown in Table 8.

Table 6. Result of Kruskall-Wallis analysis of the Classical Test Theory

Item Numbers	Asymp, Sig
1-5	0.810
6-10	0.885
11-15	0.819
16-20	0.760
21-25	0.418
26-30	0.882
31-35	0.344
36-40	0.760

Table 7. Category of power discrimination of five test packages

Discrimination Power	Package 1	Package 2	Package 3	Package 4	Package 5
Good	29 items (72.5%)	27 items (67.5%)	33 items (82.5%)	33 items (82.5%)	32 items (80%)
Poor	11 items (27.5%)	13 items (22.5%)	7 items (17.5%)	7 items (17.5%)	8 items (20%)

Table 8. Unidimension assumption test

Test Packages	KMO and Bartlett's Test		Total Variance Explained		Category
	KMO	Sig.	*Eigenvalue* Factor 1	*Eigenvalue* Factor 2	
Package 1	0.469	0.00	3.637	2.831	Multidimension
Package 2	0.513	0.00	3.807	2.223	Multidimension
Package 3	0.608	0.00	5.891	2.367	Unidimension
Package 4	0.571	0.00	5.345	2.483	Unidimension
Package 5	0.580	0.00	5.003	2.446	Unidimension

Table 8 presents that of the five packages whose unidimension assumption was analyzed, three packages are unidimensional (Package 3, Package 4, and Package 5), while two packages are multidimensional (Package 1 and Package 2). The analysis was based on the size of the sample sufficiency value (KMO) and eigenvalue. The second assumption is the local independence assumption and parameter invariance. According to Retnawati (2014, p. 7), this assumption is automatically proved after it is proved with unidimensionality.

After the assumption testing, the test item characteristic was analysed by the IRT. Testing the fitness of each item to the model followed the formula by Sumintono and Widhiarso (2015, p. 81) that an item fits to a model if the value of Outfit MNSQ is between 0.5 and 1.5. An item difficulty index can be known from the most difficult, moderate, and easiest item. An item difficulty index is categorized easy if it has the difficulty index close to -2.00. An item difficulty index is categorized moderate if its difficulty index value ranges from -1.00 to +1.00. An item difficulty index is categorized difficult if its difficulty index is close to +2.00. The result of the analysis of item characteristics based on difficulty index is shown in Table 9.

Table 9. Characteristics of items in Package 1 based on the Item Response Theory

Item Number	Model Fitness	Category	Difficulty index	Category	Category
1	1.77	Not Fit	0.390	Moderate	Poor
2	0.91	Fit	0.270	Moderate	Good
3	0.97	Fit	0.350	Moderate	Good
4	1.26	Fit	0.270	Moderate	Good
5	0.96	Fit	0.190	Moderate	Good
6	1.16	Fit	0.200	Moderate	Good
7	0.69	Fit	0.460	Moderate	Good
8	0.91	Fit	0.210	Moderate	Good
9	0.75	Fit	0.370	Moderate	Good
10	0.82	Fit	0.200	Moderate	Good
11	0.88	Fit	0.210	Moderate	Good
12	0.94	Fit	0.190	Moderate	Good
13	1.24	Fit	0.240	Moderate	Good
14	0.64	Fit	0.420	Moderate	Good
15	1.03	Fit	0.190	Moderate	Good
16	0.87	Fit	0.210	Moderate	Good
17	0.62	Fit	0.320	Moderate	Good
18	0.99	Fit	0.240	Moderate	Good
19	1,01	Fit	0.190	Moderate	Good
20	1.25	Fit	0.210	Moderate	Good
21	0.96	Fit	0.460	Moderate	Good
22	0.91	Fit	0.200	Moderate	Good
23	1.39	Fit	0.300	Moderate	Good
24	1.13	Fit	0.200	Moderate	Good
25	0.97	Fit	0.210	Moderate	Good
26	0.94	Fit	0.190	Moderate	Good
27	0.81	Fit	0.220	Moderate	Good
28	1.07	Fit	0.340	Moderate	Good
29	0.94	Fit	0.210	Moderate	Good
30	1.16	Fit	0.190	Moderate	Good
31	0.80	Fit	0.200	Moderate	Good
32	0.86	Fit	0.300	Moderate	Good
33	0.83	Fit	0.190	Moderate	Good
34	1.00	Fit	0.200	Moderate	Good
35	1.13	Fit	0.210	Moderate	Good
36	0.88	Fit	0.280	Moderate	Good
37	1.28	Fit	0.190	Moderate	Good
38	0.96	Fit	0.200	Moderate	Good
39	1.00	Fit	0.200	Moderate	Good
40	1.04	Fit	0.190	Moderate	Good

Table 9 shows that, in terms of good criteria items, 39 items fit, and one item does not fit to Rasch model because it is outside the stated OUTFIT MNSQ range. Furthermore, in terms of the item difficulty index, all items fall into the moderate category, and therefore it can be concluded that only one of the 40 items is not good. Later, based on the result of the analysis of Package 1, the analysis of the difficulty index of the items in the other test packages was conducted. The item analysis using the Item Response Theory of five test packages resulted in the value of parameter of the difficulty index of each item as shown in Table 10.

Table 10 shows the difficulty value of each test item in five test packages after the item is suited to the items in Package 1. Baker (2001, p. 11) divides difficulty indices of items according to the IRT into five categories: very easy, easy, moderate, difficult, and very difficult. An item is said to be very easy if its difficulty index value is lower than -2.00. An item is categorized easy if it has the difficulty index value close to -2.00. An item is categorized moderate if it has the difficulty index value

Table 10. Difficulty index of five test packages based on the Item Response Theory

Item Number	Difficulty Index				
	Package 1	Package 2	Package 3	Package 4	Package 5
1	0.390	0.420	0.200	0.180	0.390
2	0.270	0.250	0.190	0.300	0.250
3	0.350	0.310	0.220	0.180	0.220
4	0.270	0.280	0.300	0.460	0.470
5	0.190	0.190	0.190	0.180	0.170
6	0.200	0.190	0.220	0.170	0.180
7	0.460	0.310	0.240	0.470	0.460
8	0.210	0.200	0.250	0.220	0.180
9	0.370	0.310	0.210	0.250	0.300
10	0.200	0.200	0.210	0.390	0.180
11	0.210	0.200	0.290	0.180	0.180
12	0.190	0.200	0.280	0.180	0.180
13	0.240	0.210	0.220	0.220	0.190
14	0.420	0.260	0.360	0.250	0.350
15	0.190	0.190	0.190	0.190	0.170
16	0.210	0.220	0.200	0.170	0.190
17	0.320	0.300	0.340	0.350	0.250
18	0.240	0.270	0.200	0.190	0.220
19	0.190	0.200	0.290	0.180	0.180
20	0.210	0.230	0.200	0.180	0.180
21	0.460	0.340	0.210	0.180	0.420
22	0.200	0.200	0.260	0.180	0.180
23	0.300	0.210	0.200	0.200	0.180
24	0.200	0.250	0.200	0.180	0.170
25	0.210	0.230	0.210	0.200	0.180
26	0.190	0.200	0.260	0.180	0.200
27	0.220	0.220	0.200	0.170	0.180
28	0.340	0.370	0.190	0.180	0.200
29	0.210	0.210	0.210	0.180	0.180
30	0.190	0.190	0.240	0.420	0.180
31	0.200	0.210	0.520	0.190	0.190
32	0.300	0.270	0.190	0.180	0.240
33	0.190	0.210	0.220	0.170	0.180
34	0.200	0.420	0.210	0.180	0.190
35	0.210	0.240	0.210	0.220	0.190
36	0.280	0.300	0.210	0.190	0.220
37	0.190	0.190	0.230	0.190	0.180
38	0.200	0.200	0.270	0.180	0.170
39	0.200	0.210	0.220	0.240	0.180
40	0.190	0.200	0.190	0.190	0.190
Average	**0.250**	**0.245**	**0.236**	**0.222**	**0.222**

ranging from -1.00 to +1.00. An item is categorized difficult if it has the difficulty index value close to +2.00, and categorized as very difficult if the difficulty index value is higher than +2.00. Based on the result of the analysis using the Item Response Theory, all items in Package 1, Package 2, Package 3, Package 4 and Package 5 have the difficulty index in a good category. This is in line with Table 8 which shows that all difficulty indexes of the items range from higher than -1.00 to lower than 1.00, which means that all items have the difficulty index in the moderate category.

In addition to showing item characteristics based on difficulty index according to the IRT, Table 10 also shows the average difficulty index of 40 test items in five test packages. Table 10 shows that the average difficulty index of the test items in Package 1 is 0.250, in Package 2 it is 0.245, in Package 3 it is 0.236, in Package 4 it is 0.222, and in Package 5 sit is 0.222. Table 10 also shows that all items in five packages have the difficulty index which is not very different from each other. Based on the result of the analysis using the Classical Test Theory, a test was done to see the significance of the differences in item difficulty index among the randomized test packages. The test was conducted using Kruskall-Wallis analysis. The summary of the analysis result is presented in Table 11.

Table 11. The result of the test using Kruskall-Wallis of the Classical Test Theory

Item Number	Asymp. Sig
1-5	0.591
6-10	0.795
11-15	0.178
16-20	0.222
21-25	0.063
26-30	0.094
31-35	0.054
36-40	0.110

Table 11 shows the value of Asymp, Sig of all items whose difference among Package 1, Package 2, Package 3, Package 4, and Package 5 is above 0.05. This means that there is no difference in the difficulty index of the five test packages. Therefore, there is no effect of item number randomization on the item difficulty index.

Discussion

Mathematics is one of the school subjects which is tested in junior high school national examination. Hamdi, Kartowagiran, and Haryanto (2018) believe that students' mathematics competence can be used to solve varieties of problems and difficulties they face in learning various sciences, especially natural science. This fact forms the basis for the importance of mathematics, so that it becomes one of the school subjects examined in the national examination. The mathematics test in the national examination consists of a number of parallel test packages. The packages are constructed with the same items but with randomized item numbers and alternative answers in order to distinguish one package from the others. The use of parallel test packages is expected to prevent students from cheating, so that their real mastery can be known. Unparallel tests may result in error of measurement, that is, the result of the test does not show the real competence mastery of the students (Purnama, 2017). This research is conducted by analysing five test packages which are different based on the item randomization in order to prove whether being randomized the test packages are really parallel.

Whether or not a test is of good quality can be seen in the difficulty index of each item. A test item is said to be good if it is neither too difficult nor too easy, or in other words, the difficulty index is moderate. The item difficulty index is usually related to the aim of the test (Mehrens & Lehmann, 1973, p. 195). This research applies the Classical Test Theory and the Item Response Theory approaches in the analysis of test item difficulty index. The Classical Test Theory approach is a very simple approach and easy to understand in analyzing test items empirically (Gűler, Uyanik, & Teker, 2014), while the Item Response Theory approach is used to cover the weaknesses of the Classical Test Theory approach.

Before a further analysis was conducted to find out whether a test remained parallel after its item numbers were randomized, the quality/characteristic function of the items in Package 1 was analysed, because Package 1 is

the original test package as the reference for the analysis of the other four test packages. Putro (2013) states that good test items have to meet at least three requirements, i.e. item difficulty index, discrimination power, and well-functioned distractors. The result of the analysis using the Classical Test Theory shows that in general Package 1 is in a poor category. This can be seen in the difficulty index, discrimination power, and the functioning of the distractors. Viewed from the value of the difficulty index, it is very obvious that there are still many items in the easy category, and thus the students can answer correctly.

The result of the analysis of the five test packages using the Classical Test Theory approach shows that, in terms of the difficulty index, out of the 40 test items in five test packages, 5% to 7.5% of the items are difficult items, 35% to 62.5% of the items are moderate or good items, and 30% to 60% of the items are easy items. Viewed from the average of the item difficulty index as shown in Table 4, all of the five test packages have the average difficulty index categorized moderate or good. The value of the item difficulty index of the five test packages lies between 0.102 and 0.968. The higher the difficulty index, the easier the test item will be, and vice versa, the lower the item difficulty index, the more difficult the item will be (Bichi, 2016). This is in line with Allen and Yen (1979) who state that in test item measurement, the item difficulty index is related to the percentage of the examinees who can do the test correctly. Difficulty index is the proportion of the number of test takers who answer a particular question correctly, the proportion of all test takers.

Based on the Classical Test Theory, it is known that there has been a shift in the category of the difficulty index of some items in Package 2, Package 3, Package 4, and Package 5 compared to that of the items in Package 1. For example, test item 1 in Package 1 is in the easy category, in Package 3 and Package 4 it is in the moderate category. Another example is that test item 13 in Package 1 is in the easy category, but in Package 2 it is in the moderate category. Overall, the percentage of the shift of the category of the difficulty index of Package 2 is 22.5%, Package 3 is 60%, Package 4 is 50% and Package 5 is 37.5%. This is due to the weakness of the result of the item analysis using the Classical Test Theory approach, i.e. the size of the item characteristics (in this case the difficulty index) depends on the distribution of the competence of the test takers in the sample that is used (Awopeju & Afolabi, 2016). In line with this opinion, Zaman, Kashmiri, Mubarak, and Ali (2008) add that the comparison of test result of different test takers is one of the weaknesses of the Classical Test Theory which is worth noting, because test takers must do the items which are the same or really parallel. It is one of these weaknesses that necessitate the IRT to come into use.

In the IRT, the first thing to see is the assumption test. The unidimension assumption testing of the five test packages must first see the sufficiency of the sample. Research findings show that the value of KMO-MSA of Package 1 is 0.469, Package 2 is 0.513, Package 3 is 0.608, Package 4 is 0.571, and Package 5 is 0.580. According to Field (2009), the value of KMO-MSA is considered sufficient if it is above 0.5. From this result, it can be concluded that four packages have sufficient sample, i.e. Package 2, Package 3, Package 4, and Package 5, because the value of KMO-MSA >0.5. The result of the significance analysis using Barlett's Test of Sphericity shows that each of the five test packages is at the significance level of 0.000. Therefore, the requirement is met because the significance level is below 0.05.

There are a number of ways to interpret the sufficiency of unidimension assumption. One of the ways is by looking at the contribution of the first eigen value to test variance. The result of the above analysis shows that three test packages have dominant factors whose value is more than twice as much as the second factor, i.e. Package 3 with 5.891 which is higher than the *eigenvalue* of the second factor of 2.367. Package 4 with 5.345 higher than the eigenvalue of the second factor of 2.483, and Package 5 with 5.003 345 higher than the eigenvalue of the second factor of 2.446, where the first factor is the most dominant factor. In the factor analysis, the

first eigenvalue should have the highest value (dominant) compared to the second, third, and so forth eigenvalue. This is because the size of the variance is directly proportional with the size of eigenvalue (Field, 2009, p. 652; Johnson & Wichern, 2002, p. 441), and therefore, it can be concluded that the first factor in the factor analysis contributes the most compared to the other factors, and thus the unidimensionality assumption is met.

Difficulty index (b) which lies between the range of -2 and 2 is good (Surya & Aman, 2016). The result of the analysis using the IRT approach shows that the five test packages have the difficulty index ranging from 0.170 to 0.470. The value of the difficulty index shows that all of the test items are in the moderate category, which lies between -1.00 and +1.00 (Sumintono & Widhiarso, 2015). It means that based on the result of the analysis using the IRT, all test items in the five test packages have the same characteristics.

The analysis of the characteristics of the item difficulty index was then followed by Kruskall-Wallis analysis to reveal the effect of the randomization of the item numbers and alternative answers on the item difficulty index. The Kruskall-Wallis analysis was conducted using the value of the difficulty index obtained using the Classical Test Theory and Item Response Theory approaches. The result of the analysis using the Classical Test Theory approach shows that the randomization of the item numbers and alternative answers does not affect the item difficulty index as shown by the value of Asymp, Sig above 0.05. This result is in line with the finding of the research by Santoso (2013) which states that the estimation of the competence and length of the test with randomized design is not significantly different from the test which was not randomized. The research finding applying the IRT approach shows that there is a difference in the difficulty index of the test items in the five test packages after the randomization.

In relation to the case of the Classical Test Theory approach, the absence of the effect of the randomization of the item numbers and alternative answers may result from Package 1 which is the original test package not having undergone any randomization. Package 1 has the characteristics which tend to be poor. Viewed from its difficulty index, more than 50% of the items are easy items which make most students, those with high competence and those with low competence, can answer questions correctly. It means that the test cannot distinguish students with high competence from those with low competence. A test that tends to be easy for students will not show any effect of randomization because they will tend to be able to do it. In addition, a test was conducted using Kruskal-Wallis test on the difficulty index using the Classical Test Theory and Item Response Theory. Package 1 which is the original test package is used to find out whether there is a difference in the difficulty index between items 1-10 and items 31-40. The result shows the Assymp, Sig value of 0.082 when using the Classical Test Theory, and the Assymp, Sig value of 0.054 when using the IRT, where the Assymp, Sig value is above 0.05. It means that in the original test package, before randomization, the values of the item difficulty index are not in a wide range. This may be the reason for the absence of the difference in the difficulty index after randomization. Further studies need to be done on the test items which have good characteristics to see whether or not there is an effect of the randomization of the item numbers and alternative answers on item difficulty index.

Conclusion

All of the five test packages have a good reliability index, lying between 0.96 and 0.97. Package 1, Package 2, and Package 3 have the reliability index of 0.96, while Package 4 and Package 5 have the reliability index of 0.97. It can be concluded that based on the value of the reliability index, the five test packages have equal reliability.

Based on the result of the analysis using the Classical Test Theory, viewed from the average value of the difficulty index, all five test packages have the average difficulty index ranging from 0.102 to 0.968. The result of Kruskall-Wallis analysis of the five test packages shows that there is no difference in the difficulty index of the items in Package 1,

Package 2, Package 3, Package 4 and Package 5. Thus, the randomization of the item numbers and alternative answers has no effect on the item difficulty index.

The analysis of the test items using the Item Response Theory shows that the average value of difficulty index of the five test packages ranges from 0.170 to 0.470. The result of the analysis of the difficulty index of the items in the five test packages shows that there is no difference in the difficulty felt by the students doing Package 1, Package 2, Package 3, Package 4, and Package 5. This means that the randomization of item numbers has no effect on the item difficulty index, which means that constructing parallel tests by randomizing the item numbers and alternative answers is good to do, and this research has proved that applying this method will result in parallel tests.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Los Angeles, CA: Wadsworth.

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of Classical Test Theory and Item Response Theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal, ESJ*, *12*(28), 263–284. https://doi.org/10.19044/esj.2016.v12n28p263

Azwar, S. (2013). *Reliabilitas dan validitas* (4th ed.). Yogyakarta: Pustaka Pelajar.

Azwar, S. (2015). *Reliabilitas dan validitas*. Yogyakarta: Pustaka Pelajar.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Bichi, A. A. (2016). Classical Test Theory: An introduction to linear modeling approach to test and item analysis. *International Journal for Social Studies*, *2*(9), 27–33. https://doi.org/10.26643/ijss.v2i9. 6690

Center for Educational Assessment. (2014). *Laporan pengolahan Ujian Nasional tahun ajaran 2014/2015* (Unpublished). Jakarta: Center for Educational Assessment of Republic of Indonesia.

Fernandes, H. J. X. (1984). *Testing and measurement*. Jakarta: National Education Planning, Evaluation, and Curriculum Development.

Field, A. (2009). *Discovering statistics using SPSS* (3rd 3d.). London: Sage Publications.

Gǖler, N., Uyanik, G. K., & Teker, G. T. (2014). Comparison of Classical Test Theory and Item Response Theory in terms of item parameters. *European Journal of Research on Education*, *2*(1), 1–6.

Hamdi, S., Kartowagiran, B., & Haryanto, H. (2018). Developing a testlet model for mathematics at elementary level. *International Journal of Instruction*, *11*(3), 375–390. https://doi.org/10.12973/iji.2018.11326a

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Kronmüller, K.-T., Saha, R., Kratz, B., Karr, M., Hunt, A., Mundt, C., & Backenstrass, M. (2008). Reliability and validity of the knowledge about depression and mania inventory. *Psychopathology*, *41*(2), 69–76. https://doi.org/10.1159/000111550

Law No. 14 of 2005 of Republic of Indonesia about Teachers and Lecturers. , (2005).

Mardapi, D. (2014). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Litera.

Mehrens, W. A., & Lehmann, J. L. (1973). *Measurement and evaluation in education and psychology*. New York, NY: Holt, Rinehart, and Winston.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.

Naga, D. S. (1992). *Pengantar teori sekor pada pengukuran pendidikan*. Jakarta: Gunadarma.

Purnama, D. N. (2017). Characteristics and equation of accounting vocational theory trial test items for vocational high schools by subject-matter teachers' forum. *REiD (Research and Evaluation in Education)*, *3*(2), 152–162. https://doi.org/10.21831/reid.v3i2.18121

Putro, N. H. P. S. (2013). Karakteristik butir soal ulangan kenaikan kelas sebagai persiapan bank soal Bahasa Inggris. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *15*(1), 92–114. https://doi.org/10.21831/pep.v15i1.1089

Rasyid, H., & Mansur, M. (2008). *Penilaian hasil belajar*. Bandung: CV Wacana Prima.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207–230. https://doi.org/10.3102/10769986004003207

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson.

Rohmawati, R. (Ed.). (2013). Kurikulum 2013, 87 persen guru kesulitan cara penilaian. Retrieved January 6, 2018, from https://unnes.ac.id/berita/87-persen-guru-kesulitan-soal-penilaian-kurikulum-2013.html

Sanjaya, W. (2010). *Kurikulum dan pembelajaran*. Jakarta: Kencana.

Santoso, A. (2013). Pemilihan butir alternatif pada tes adaptif untuk peningkatan keamanan tes. *Jurnal Kependidikan: Penelitian Inovasi Pembelajaran*, *43*(1), 1–8. https://doi.org/10.21831/jk.v43i1.1953

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Cimahi: Trim Komunikata.

Surya, A., & Aman, A. (2016). Developing formative authentic assessment instruments based on learning trajectory for elementary school. *REiD (Research and Evaluation in Education)*, *2*(1), 13–24. https://doi.org/10.21831/reid.v2i1.6540

Werheid, K., Hoppe, C., Thone, A., Muller, U., Mungersdorf, M., & von Cramon, D. Y. (2002). The adaptive digit ordering test clinical application, reliability, and validity of a verbal working memory test. *Archives of Clinical Neuropsychology*, *17*(6), 547–565. https://doi.org/10.1093/arclin/17.6.547

Zaman, A., Kashmiri, A.-U.-R., Mubarak, M., & Ali, A. (2008). Students ranking, based on their abilities on objective type test: Comparison of CTT and IRT. *Edu-Com International Conference*, 591–599. Retrieved from https://ro.ecu.edu.au/ceducom/52/